AHMED, ALAA HASSAN, M.S. Fusing Uncertain Data with Probabilities. (2016)
Directed by Dr. Fereidoon Sadri. 86pp.

Discovering the correct data among the uncertain and possibly conflicting mined data is the main goal of data fusion. The recent research in fusing uncertain data shows that taking source confidence into account helps to achieve this goal because the sources have different degree of accuracy. Thus, understanding different modern fusing techniques and using different data sets can be useful to research community.

Previous work has fused uncertain data with and without considering correlation between the sources by using training datasets [5]. In our proposed research, we extended this work by calculating the initial probability which is given by the sources that provide the information and then calculating the final probability for the given data. In our work there is no need to training set in which the algorithm can work with different type of uncertain datasets. Also, we present a method to calculate the threshold of the given dataset; and we did our experiments by using two types of datasets; one type contains intentional false and other random false.

FUSING UNCERTAIN DATA WITH PROBABILITIES

by

Alaa Hassan Ahmed

A Thesis Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Greensboro
2016

Approved by

_____
Committee Chair

All the praise and thanks be to Allah for helping me to complete my master's degree successfully.

To my great country, Iraq, thank you for giving me the opportunity to get my degree.

To my lovely family especially my mother and father, thanks for supporting me.

To my brother and sisters, thank you for being always with me.

To my fiancé, thank you for all of your encouragement.

To my friends, thanks for everything.

APPROVAL PAGE

This thesis written by Alaa Hassan Ahmed has been approved by the following committee of the faculty of the Graduate School at The University of North Carolina at Greensboro.

Committee Chair _____
Fereidoon Sadri

Committee Members _____
Jing Deng

_____
Lixin Fu

April 21, 2016
_____
Date of Acceptance by Committee

April 21, 2016
_____
Date of Final Oral Examination

ACKNOWLEDGMENTS

I would like to dedicate this page to share my gratitude for Dr. Fereidoon Sadri. Truly, words cannot express how grateful I am for having him as my advisor. With all sincerity, I would like to thank him for his guidance, wisdom, and care. I appreciate the time he spent helping me deeply understand the material and tasks in order for me to be sufficient at what I am doing. I would also like to take this opportunity to thank Dr. Jing Deng and Dr. Lixin Fu for their valuable guidance and feedback.

TABLE OF CONTENTS

Page

LIST OF TABLES vii

LIST OF FIGURES ix

CHAPTER

I. INTRODUCTION 1

1.1 Introduction 1
1.2 Uncertain Data 2
1.3 Representation Model 3
1.3.1 Possible Worlds 3
1.3.2 Probabilistic Relation Model (pr-relation) 5
1.3.3 Extended Probabilistic Relation (epr-relation) 5
1.4 Integrating Uncertain Data 6
1.5 Fusing Uncertain Data 7
1.6 Probabilistic Database 9
1.7 Contribution from This Work 10

II. INTEGRATING UNCERTAIN DATA 12

2.1 Possible World without Assigned Probabilities 12
2.2 Possible World with Assigned Probabilities 15
2.3 Using Probabilities with Data Fusion 19

III. FUSING UNCERTAIN DATA 20

3.1 Data Fusion 20
3.2 Distinguish between True and False Data 21
3.3 Fusing Data by Considering Independency between Sources 22
3.3.1 Quality of the Source 24
3.3.2 Estimate the Probability for Each Triple 28
3.4 Fusing Correlated Sources 31
3.4.1 Advantages of Correlation between Sources 33
3.4.2 Positive Correlation and Negative Correlation 35
3.4.3 Exact Solution 35
3.4.4 Aggressive Approximation 37
3.4.5 The Relation between Sources and Triples 38

IV. PROBABILITIES FOR UNCERTAIN DATA FUSING 40

4.1 Calculating Probabilities in Data Fusion 40
4.1.1 Considering the Independencies between the Sources 40
4.1.2 Considering the Correlation between the Sources. 42
4.1.2.1 Exact Solution 43
4.1.2.2 Aggressive Approximation 44
4.2 Example for Counting the Probabilities 45
4.2.1 Considering the Independencies between Given Sources 46
4.2.2 Considering the Correlation between the Sources 48
4.2.2.1 Exact Solution 48
4.2.2.2 Aggressive Approximation 49

V. DATA SETS 50

5.1 Training Data Set 50
5.2 Countries and Capitals Data Set (Intentional False) 52
5.3 Countries and Capitals Data Set (Random False) 54
5.4 Toy Example Data Set 57

VI. RESULTS AND FINDINGS 59

6.1 Results 59
6.2 Independence Case 60
6.3 Correlation Case 75
6.3.1 Exact Solution 75
6.3.2 Aggressive Approximation 78
6.4 Comparing between the Results 80

VII. CONCLUSION AND FUTURE WORK 82

7.1 Conclusion 82

REFERENCES 84

LIST OF TABLES

Page

Table 1. Possible World Model for the Temperature of New York City 4

Table 2. Probabilistic Relation Model for the Temperature of New York City 5

Table 3. Extended Probabilistic Relation Model for Representing Uncertain Data 6

Table 4. The Possible World of the Two Sources S1 and S2 14

Table 5. The Possible World of the Source S1 17

Table 6. The Possible World of the Source S2 17

Table 7. Summary of Notations Used in Fusing Data 25

Table 8. Precision and Recall for Each Source 27

Table 9. False Positive Rate of the Five Sources 30

Table 10. False Positive Rate for Selected Subset of Sources 33

Table 11. Mini Dataset with 5 Sources and 9 Triples 46

Table 12. Precision, Recall, and False Positive Rate for the 5 Sources 47

Table 13. Recall and False Positive Rate for the Selected Subset of Sources 48

Table 14. Data Extracted by Five Different Extractors from the Wikipedia Page for Barack Obama 52

Table 15. The First 8 Triples of Countries and Capitals (Intentional False) Dataset 54

Table 16. The First 9 Triples of Countries and Capitals (Random False) Dataset 56

Table 17. Data Extracted by Five Different Extractors from the Wikipedia Page for Barack Obama with Probabilities Added to it 58

Table 18. The Average CPU Time of Five Times Running and Sources p, r, and q by Considering Independency between the Sources Using Different Number of Sources 61

Table 19. The Average CPU Time of Five Times Running and Sources p, r, and q for Different Number of Sources 63

Table 20. The Average CPU Time of Five Times Running and Sources p, r, and q for Different Number of Sources 64

Table 21. The Average CPU Time of Five Times Running and Sources p, r, and q for Different Number of Sources 66

Table 22. The Average CPU Time of Five Times Running and Sources p, r, and q for Different Number of Sources 67

Table 23. The Average CPU Time of Five Times Running and Sources p, r, and q for Different Number of Sources 69

Table 24. The Average CPU Time of Five Times Running and Sources p, r, and q for Different Number of Sources 70

Table 25. The Average CPU Time of Five Times Running and Sources p, r, and q for Different Number of Sources 72

Table 26. The Average CPU Time of Five Times Running and Sources p, r, and q for Different Number of Sources 73

Table 27. The Average CPU Time of Five Times Running and Sources p, r, and q for Different Number of Sources 74

Table 28. CPU time by Considering Correlation between the Sources Using Different Number of Sources 76

Table 29. CPU time by Considering Correlation between the Sources Using Different Number of Sources 77

Table 30. CPU time by Considering Correlation between the Sources Using Different Number of Sources 79

LIST OF FIGURES

Page

Figure 1. Consistency Graph of the Two Sources S1 and S2 18

Figure 2. Precision and Recall 26

Figure 3. The Wikipedia Page of Barak Obama and Five Extractors that Extract Knowledge from it 51

Figure 4. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 10% Random Errors 61

Figure 5. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 20% Random Errors 62

Figure 6. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 30% Random Errors 64

Figure 7. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 40% Random Errors 65

Figure 8. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 50% Random Errors 67

Figure 9. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 10% Intentional Errors 68

Figure 10. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 20% Intentional Errors 70

Figure 11. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 30% Intentional Errors 71

Figure 12. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 40% Intentional Errors 73

Figure13. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 50% Intentional Errors 74

Figure 14. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 10% Random Errors 76

Figure 15. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 20% Random Errors 77

Figure 16. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 10% Random Errors 79

Figure 17. The Precision of 5 Sources with Using the Probabilistic Theory Approach to Compute Initial Probabilities 80

Figure 18. The Precision of 5 Sources with Using an Ad-hoc to Compute Initial Probabilities 80

CHAPTER I

INTRODUCTION

1.1 Introduction

In recent years, uncertain data fusing, integrating, modeling and managing has received significant interest and became one of the most important subjects to search about [1] [2] [3] [14] [15]. Fusing uncertain data from multiple sources provides a unified view of data; and the fused data turns to be more trustworthy, meaningful and more accurate than the data provided by individual source. Also, it helps to make different data sources work together which helps and frees the user from tedious task of finding relevant source. Data fusion is a form of information integration where large amounts of data mined from sources such as web sites, Twitter feeds, Facebook postings, blogs, email messages, and the like are integrated. Such data is inherently uncertain and unreliable. The sources have different degrees of accuracy and the data mining process itself incurs additional uncertainty. The main goal of data fusion is to discover the correct data among the uncertain and possibly conflicting mined data. To illustrate more, suppose a data-integration system provides information about movies from different data sources; and suppose that we want to find a certain movie and the reviews about it. No one of these sources can answer this query in isolation ,which makes the integration imperative in order to answer this type of queries [9].

Moreover, these processes became very important in many life aspects such as theoretical, commercial and scientific; and it has a huge benefit in providing an accurate data for many real world application which faces with uncertainty. Within the growing number of such applications, it becomes an essential purpose to retrieve efficient and scalable information in answering user's complex queries. Since there are still a lot of data sources that provides information with varying degree of certainty fusing, integrating and answering some types of queries on uncertain data continue to be a challenging area to research [11].

1.2 Uncertain Data

Uncertain data also called as symbolic data which contains noise that keeps it out of the truth. Uncertain data generates by many new modern applications such as information extraction, data cleaning, de-duplication and many other applications; and in some applications uncertain data sets are inherent such as environmental surveillance, market analysis, and quantitative economics research [11] [12] [17]. Also, a lot of technologies that collect the data in imprecise way can cause uncertainty in data in everywhere [10]. Uncertain data in those types of applications are generally can caused by being outcome of flawed data, missing knowledge, incompleteness, limitations of measuring equipment, delayed data updates, etc. In addition, there are many reasons could lead to uncertainty such as Measurement Errors, Multiple or Inconsistent Sources and Approximate Schema Mapping. Those types of unreliable data should be handled with caution in order to not cause doubtful results in integrating or fusing data [12].

Since a traditional database is not able to handle uncertain data, then there is a big need to database that can deal with data with varying degree of certainty. Therefore, a lot of studies focused on handling uncertainty in databases. The work in [1] [2] [3] explained methods to solve uncertain data by using representation model; and the best definition of uncertain database is defined in [3] as: An uncertain database U consists of a finite set of triples T (U) and a nonempty set of possible worlds PW (U) = {D1,…, Dm}, where each Di Є T (U) is a certain database.

1.3 Representation Model

The work in [1] [2] [3] represents uncertain information by using different models such as possible worlds, probabilistic relation model, and extended probabilistic relation.

1.3.1 Possible Worlds

Possible world model is a conceptual model of uncertain data. Each source in possible worlds represents as an instance being a possible state and it doesn't contain any uncertainty. To illustrate more, suppose there are two sources which provide the temperature of New York City for one day. One of them recorded as 50 and the other as 53. So, there will be 4 possibilities as its shown in table 1:

Table 1. Possible World Model for the Temperature of New York City

Triple	Temperature
1	50

Possible world D1

Triple	Temperature
1	53

Possible world D2

Triple	Temperature
1	50
2	53

Possible world D3

ϕ which means nothing

Possible world D4

As it is obvious in the previous example that the last two possible worlds D3 and D4 are wrong because one city cannot have two temperatures; also it cannot be without temperature. Thus, the answer is 50 or 53. Still there are some difficulties in dealing with the possible worlds because the number of possible worlds increases as the number of uncertain data increases; and that what need more time to spend in order to solve them.

Note that, uncertain database that contains fewer possible worlds contains more data than uncertain database that contains more possible worlds and this if both the databases have the same set of triples [3].

1.3.2 Probabilistic Relation Model (pr-relation)

The differences between the pr-relation and possible worlds is that pr-relation just uses one schema with new attribute called as event attribute (E) but in possible worlds case, it uses multiple schemas depending on the number of possibilities that it can be get from given uncertain data. Event variable are expressed by using Boolean variables, true or false, and it can connect by logical symbols to make it more complex. For example, the temperature example can be expressed by using x and y for the two possibilities; and for the simplicity, we can use ¬x instead of y for the second probabilities. The table below shows the pr-relation for the temperature example.

Table 2. Probabilistic Relation Model for the Temperature of New York City

Triple	Temperature	Events
1	50	X
2	53	¬ x

1.3.3 Extended Probabilistic Relation (epr-relation)

Pr-relation has been extended in [2] which adds event constraints and it called as extended probabilistic relation (epr-relation). To illustrate it, consider the example from [2], Andy and Jane are talking about a student called John. Andy says “I am taking CS100, CS101, and CS102 and John is in either CS100 or CS101 but not in both”. Jane says “I am taking CS101 and CS102 and John is in one of them, but not in both." The following table shows the epr- relation of the example.

Table 3. Extended Probabilistic Relation Model for Representing Uncertain Data.

Triple	Temperature	Events
1	John csc100	X
2	John csc101	¬ x
3	John csc102	¬ y
	¬x ≡ y ¬y≡ false	

The constraint in the above example used to show that ¬y ≡ false because Andy said that John is not in csc102. Also, ¬x ≡ y shows that both ¬x and y are the same.

1.4 Integrating Uncertain Data

Nowadays, a lot of companies have different number of branches, and they need unified view of their data. For that reason, they need a powerful application that can merge all of their data and clear it form uncertainty. This operation called as integration. Data integration is very necessary and important because integrating uncertain data from multiple sources can fix some uncertainty and it overcomes conflict data, which yields more accurate information than any of individual sources [19]. This means, the information that is produced by integration is trustable. As an example taken from [3] of integrating uncertain data, if there is two sensors and one of them reports that an object is either in location A or in location B, and a the other reports that it is either in location B or in location C, by integrating the sensor reports we conclude that the object is in location B.

1.5 Fusing Uncertain Data

In recent database research literature, data fusion refers to the integration of massive data mined from sources such as web sites, Twitter feeds, Facebook postings, blogs, and email messages [5] [15][20][21][22][23]. The state of the art data fusion research assumes a simple uncertainty model, where data (e.g., each triple) is independent. On the other hand, sophisticated integration techniques, often based on the Bayesian analysis and Bayesian networks, and are employed in the integration process. Given that sources may provide erroneous and contradictory data, the primary goal of data fusion is to determine which data are true in reality, and which are false. Alternatively, many approaches provide a measure, or probability, of correctness for each datum.

Early approaches to data fusion were based on the simple voting or counting approaches: Data provided by the majority of sources or a number of sources exceeding a given threshold number are considering true. More, recent approaches attempt at obtaining higher accuracy by estimating two sets of unknown parameters: A measure of being correct for each source (often called trustworthiness or source quality) and probability of being correct for each fact (often called truthfulness of confidence). Each set of parameters is dependent and computed using the other set. Most approaches iteratively estimate each set using the other until a stable solution) is reached [5] [15] [21] [22] [23].

To further improve the accuracy, correlation among sources can be taken into account. A widespread form of correlation is when a source copies material from another.

However, other types of dependence are also possible, such as positive and negative correlations [5]. Truthfulness, of a fact is dependent on the trustworthiness of sources providing it, but source correlations also impact truthfulness. For example, if some of the sources confirming a fact copy from others in the set, then their impact should be discounted. It has been shown that these techniques increasingly improve the accuracy of predictions. In chapter III, we provide a review of a recent work on data fusion [5]. They use Bayesian analysis to derive truthfulness of facts using trustworthiness of sources. The same parameters have been used in [23] but their approach uses Bayesian networks and is quite different from [5]. The importance of fusing these two factors to model quality of sources is they take into both the positive contribution, when a source confirms a fact, and the negative correlation, when a source rejects a fact, or when it confirms a contradictory fact.

Further, the work in [5] takes into consideration correlation among sources. Correlations are modeled by sources' joint recall and joint false positive rate. They demonstrate, through intuitive examples, the importance of source correlation on the accuracy of the predication for different correlation types, copying overlap on true facts, overlap on false facts, and complementary sources. An exact computation of truthfulness is computationally expensive (exponential in the number of sources, which can be very large). Authors present approximation approaches to remedy this difficulty. Their "aggressive approximation" has linear time complexity in the number of sources, while the "elastic approximation" can be used to improve the accuracy (over aggressive approximation) incrementally until a desired accuracy has been reached. Their use of

training datasets for the computation of some of the parameters, while having the advantages of producing more accurate results especially in low quality datasets, has the downside that training sets may be hard to obtain or labor intensive for some applications.

1.6 Probabilistic Database

Probabilistic data base is the database that some of its data's correctness and value are uncertain and known only with some probability [17]. Dealing with that type of databases can cause absolute uncertain results which are not desired. To get rid of the uncertainty that probabilistic database contains, is to fuse or integrate its data which produce more accurate results.

Probabilistic database became very important since it deals with uncertainty and the normal databases are deterministic and not able to deal with uncertainty. Therefore, a lot of studies applications developed recently to deal with uncertain database and their goal are to clean its uncertainty [16]. Probabilistic database has three types of uncertainty:

- Tuple-level uncertainty
- Attribute-level uncertainty
- Correlated- level uncertainty

In our work, we are considering attribute- level uncertainty and we use fusing algorithm to find each triple's probability.

1.7 Contributions from This Work

The work of [5] has developed algorithm to fuse data from uncertain data without using known associated probabilities. In contrast, the work of [1] has integrated uncertain data with and without known associated probabilities. We extend the work of [5] by using probabilities associated to uncertain data in the following manner:

- We present a comprehensive review of data fusion work in [5] and we run the algorithm without taking the probabilities of the given triples.
- We present a comprehensive review of integrating uncertain data by using possible worlds as is done in [1].
- We combine the work that is done in both [1] and [5], by fusing uncertain data with assigned probabilities.
- We present a method to calculate the probabilities of each triple in uncertain database depending on the probabilities that is given by the sources that provides the triple.
- We compute the correctness value of each of triple using fusion algorithm.
    - By considering independency between sources.
    - By considering correlation between sources.
- We compare the results with and without using probabilities associated to uncertain data and we showed the differences.
- We did a comprehensive evaluation of our work by using multiple datasets without assigned probabilities, and multiple datasets with assigned probabilities.

- To evaluate the fusing techniques with many datasets, we present two methods to create datasets with n number of sources and m numbers of triples
  - By considering intentional false between the sources, copying between the given sources.
  - By considering random false between the sources, no copying between the sources.

The arrangement of this work is as follow:

In chapter II, it discusses about the previous work that done to integrate uncertain data with assigned probabilities. In chapter III, it discusses about fusing uncertain data without using probabilities. In chapter IV, we explained about our work of fusing data with assigned probabilities. In chapter V, it elucidates the data sets that we create and used in our work. In chapter VI, we present the results that we gain by running the algorithms in different cases and with different datasets. In the end, we give the conclusion of the work and future work.

CHAPTER II

INTEGRATING UNCERTAIN DATA

This chapter briefly reviews the work done so far towards using probabilities in integrating uncertain information. We summarize the work of [1] in using the possible world with and without assigned probabilities.

2.1 Possible World without Assigned Probabilities

The meaning of the possible world as it mentioned before is a conceptual model of uncertain data. A lot of studies used possible worlds to solve the uncertainty in datasets. In this section, we explain the possible worlds without assigned probabilities for the given information. As an extremely simple example to understand the possible worlds, one image database may label an image as blue or green, two possible worlds, while another source labels the same image as green or yellow, another two possible world. As a result of combining this two sources result, green may be deemed more likely than the other two colors [19]. The work of [1] explains the use of possible worlds in simple way as it's shown below:

Assume we have uncertain source U with a finite number of triples T (U), and a variable t_i assigned for each triple. The formula f that corresponds to the uncertain database U can be explained by the following steps:-

- Assume that the D_j is a database in the possible worlds of uncertain database.
- Build a formula by conjunction all variables, xi where the corresponding triple t_i is in D_j, and the conjunction of $\neg x_i$ where the corresponding triple t_i is not in D_j. Then, the formula can be expressed as:

$$f_j = \bigwedge_{ti \in Dj} xi \bigwedge_{ti \notin Dj} \neg xi$$

- Build a formula to the uncertain database U that is disjunction of possible worlds formulas of U. the formula can be expressed as:

$$f = \bigvee_{Dj \in PW(U)} fj$$

To illustrate it clearly suppose that there are two sources, two friends, giving information about one student who called John. Source S1, Andy, says “I am taking CS100 and CS101 and John is in one of them, but I am not sure which one.” That means John is taking either CS100 or CS 101 (but not both). Source S2, Jane says “I am taking CS101 and CSC102 and John is in one of them, but I am not sure which one.” That means John is taking either CS101 or CSC102 (but not the both). Table 4 shows the possible worlds of the given example.

Table 4. The Possible World of the Two Sources S1 and S2.

S1

D1

Student	Course
John	CS101

D2

Student	Course
John	CS102

S2

D3

Student	Course
John	CS101

D4

Student	Course
John	CS102

Let variable x_1 and x_2 correspond to each of (John, CS100) and (John, CS101) triples respectively. Then the formula for the first possible world, second possible world, and the database are, respectively:-

$$x_1 \wedge \neg x_2,\ \neg x_1 \wedge x_2, \text{ and } (x_1 \wedge \neg x_2) \vee (\neg x_1 \wedge x_2).$$

Let x_2 and x_3 correspond to (John, CS101) and (John, CS102) respectively. Then the formula for the third possible world, forth possible world and the uncertain database are, respectively:-

$$x2 \wedge \neg x3,\ \neg x2 \wedge x3,\ (x2 \wedge \neg x3)\ V\ (\neg x2 \wedge x3)$$

Now, let's find the integration between the two uncertain databases by summing them logically as its shown next:-

$$((x_1 \wedge \neg x_2) \vee (\neg x_1 \wedge x_2)) \wedge ((x_2 \wedge \neg x_3) \vee (\neg x_2 \wedge x_3))$$

To simplify this Boolean expression more, we get:

$$(\neg x1 \wedge x2 \wedge \neg x3) \vee (x1 \wedge \neg x2 \wedge x3)$$

This result shows that the student, John, is either in CS101 or in both CS100 and CS102.

2.2 Possible Worlds with Assigned Probabilities

The work in [1] extended the possible world model by adding probabilities assigned to each possible world in uncertain databases. A probabilistic uncertain database U consists of a finite set of triplets T(U), and a nonempty set of possible worlds PW(U) ={D1, . . . , Dm}.

Note that each Di ⊆ T (U) is a certain database with a probability pi, $0 \leq pi \leq$ and $\sum_{i=1}^{m} pi = 1$.

Since they are using one more attributes which is probabilities, then we can use e_i to represents the event where the value of the uncertain database U is equal to D_i. Thus, the probability of e_i, P (e_i) = pi. Depending on that, the following observation has been made:-

- A possible world is an exclusive of other possible worlds of the same source.

- In integrating set of sources S_1, S_2… S_n with uncertain databases U_1… U_n, the each possible world corresponds to the conjunction of possible worlds, one from each source.
- In integrating different sources, two worlds may be inconsistent. Then P (e_k| e_j) = P (e_j| e_k) = 0.
- For the possible worlds that are integrating, the sum of probabilities of the possible worlds of the first source should be equal to the sum of probabilities of the possible worlds of the second source.

Therefore, the probabilities which associated with possible worlds of different sources must to satisfy certain constraints called probabilistic consistency constraints. If the consistency constraints satisfies, then in integration the probabilities of individual source by using conditional probability: P ($e_j \wedge e_k$) = P (e_j|e_k) * P (e_k). However, If e_j and e_k are inconsistent, then P (e_j|e_k) = 0. To represent the consistency constrain, they are using bi-partite graph G in [1]. Assume that we have a set S with possible worlds {D1, . . ., Dk}, and a set of S' with possible worlds {D_1', …, D_k'}. If the formulas f (Di) and f (D_j') corresponding to these worlds are mutually satisfiable, then there will be an edge between Di and D_j'. To make it clearer, consider the two sources S1 and S2 with their possible worlds shown in the table 5 and 6 respectively.

Table 5. The Possible World of the Source S1.

D1

Student	Course
John	CS100

D2

Student	Course
John	CS100
John	CS101

D3

Student	Course
John	CS101

Table 6. The Possible World of the Source S2.

D1’

Student	Course
John	CS100

D2’

Student	Course
John	CS100
John	CS201

D3’

Student	Course
John	CS201

D4’

Student	Course
John	CS201
John	CS202

Figure 1 shows the consistency graph G of the two sources S1 and S2. The graph G consists of two connected sub graphs G1 which contains D1, D2, D1’, D2’ and the second G2 consist of nodes D3, D3’, D4’.

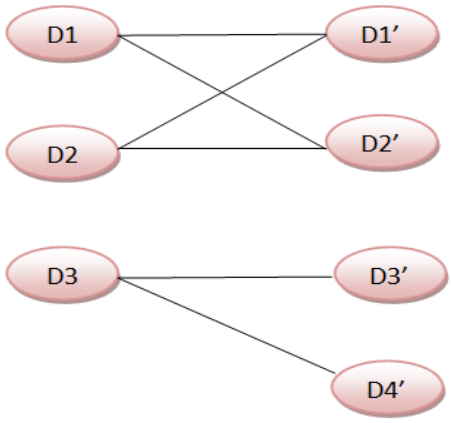

Figure 1. Consistency Graph of the Two Sources S1 and S2

Let's assume probabilities for these possible worlds of sources S1 and S2 by taking into account the consistency constraints, are P (D1) =0.3, P (D2) = 0.5, P (D3) = 0.2, P (D1') = 0.35, P (D2') = 0.45, P (D3') = 0.05, and P (D4') = 0.15.

Now by using conditional probabilities we can find the probabilities after integration:-

$$P (D3 \wedge D3') = P (D3') = 0.05,$$

$$P (D3 \wedge D4') = P (D4') = 0.15$$

We still have to find the probabilities of four remaining possible worlds in the integration. One possible way is to distribute 0.8 according to the pair wise product of probabilities of underlying possible worlds. Therefore, we will get:-

$$P (D1 \wedge D1') = 0.13125$$

$$P (D1 \wedge D2') = 0.16875$$

$$P\ (D2 \wedge D1') = 0.21875$$

$$P\ (D2 \wedge D2') = 0.28125$$

As it's clear from the results (D2 ^ D2') has the most highest probability, which means that the true answer is most highly to be John took CS100, CS101 and CS201.

2.3 Using Probabilities with Data Fusion

The work of [5] has provided algorithms to fuse data by considering independency and correlation between given sources. However, it doesn't take into account the probabilities of the triples which provided by the source. The previous works used probabilities assigned to triples in order to integrate them, but they are not highly efficient or practical in term of using high number of triples or sources. Hence, we combine the work of [1] and [5] to fuse data by using probabilities in order to get very accurate results and with large number of sources and triples. The next chapter explains how to fuse data with and without considering the correlation between the sources. After that, we present our work of fusing data with assigned probabilities considering correlation and independence between the sources.

CHAPTER III

FUSING UNCERTAIN DATA

This chapter summarizes the work done towards fusing data from uncertain databases. With this intention, we discuss data fusion and different scenarios to fuse data from uncertain database as is presented in the work of [5]. Firstly, we take into account the sources that are not correlated; and secondly we consider them correlated. In the correlated case, we discuss two methods exact solution and aggressive approximation methods to fuse data. We examined both methods and we compared them to show the difference. In this work, we extended the work done in [5] towards using probabilities assigned to each triple similar to the idea that is mentioned in the work of [1] in order to determine the probabilities of the results of fusion.

3.1 Data Fusion

Data fusion is a process that integrates different sources in order to get consistent and accurate data which is more useful [13] [14]. It has many advantages in enhancing data authenticity or availability. This subject uses widely, with different terminologist, in different science, engineering, management and many other fields [14]. In some domains like geospatial (GIS) the term fusion comes similar in meaning to the integration.

To make it more clear, we propose a review to the best definition for data fusing which is mentioned by the Joint Directors of Laboratories (JDL) workshop [4]:-

"A multi-level process dealing with the association, correlation, combination of data and information from single and multiple sources to achieve refined position, identify estimates and complete and timely assessments of situations, threats and their significance."

Many modern control system needs to have a strong fusion algorithm to combine data in a coherent manner to gain consistent and accurate results [14]. Therefore, a lot of techniques found to fuse data in order to handle uncertainty problem such Dempster-Shafer theory for evidential reasoning, fuzzy logic, neural network, Bayesian approach and statistical techniques. However, this subject still under research to find more powerful method, which can produce more accurate data and far from uncertainty.

3.2 Distinguish between True and False Data

The main idea of fusing uncertain data is to automatically distinguish the correct data from conflict data in uncertain databases for creating a cleaner set of integrated data. Majority was one of the methods to achieve this goal in naïve approach. However, naïve approach is not efficient, because the results not always correct since they are not considering copying between sources. Since the coping between sources is a common situation especially in web [15]. Therefore, a lot of researches have been done towards this idea. The work, in [5] studies fusing data and they could achieve more accurate results in distinguishing between accurate and inaccurate information. The idea behind

their success was that from given dataset and by using the steps of the algorithm, the algorithm produces the probability of the triples depending on many superior factors. So, if the probability of the triple is more than 0.5 so it's more close to be true, otherwise it's false.

3.3 Fusing Data by Considering Independency between Sources

The work of [5] fuses uncertain data taken from different sources. They proposed consecutive steps to find the true information and create a cleaner data set. We epitomize the steps of the procedure below and we give examples in each step by using the same training set that they used in [5] which we describe it clearly in chapter III (training dataset):

The data model consists of a set of sources $\{S_1, S_2, \ldots, S_n\}$, and a collection of their outputs $O = \{O_1, O_2 \ldots O_n\}$. O_i denotes to the triple provided by source $S_i \in S$; and $S_i \models t$ denotes that S_i provides t. The notation that used in the paper is shown in table 7.

The sources are deterministic, which means a source outputs a triple or not. Each source provides some information and each unit of this information called as triple. These triples are in the form of (subject, predicate, object) for example (Obama, profession, president). Each triple consider as a cell in database system in the form {row-entity, column-attribute, value} such as a row can represent Obama, column represents Profession, and value represents president.

The goal is to purge all incorrect data and gain a cleaner dataset $R = \{t: t \in O \wedge t \text{ is true}\}$. So, the triple consider being true if it's matching with the real world or it considers false. For example, (Obama, profession, president) is true; while, (Obama, surgical operation, 05/01/2011) considers false. Moreover, there are two semantic assumptions about the given data:

1. Triple independence semantic: which means the truthfulness of a certain triple is independent to other triples. As an illustration, suppose the source provides triple t1 is independent whether the same source provides triple t2 or not.
2. Open world semantic: which means that if a source provides a triple so it considers it as a true; and if it doesn't provides it then its unknown (rather than considering it false). As an illustration, suppose that source S1 provides triple t1 but not t2. So, it considers t1 as true but it doesn't know if triple t2 is provided or not and it doesn't consider it false.

These two assumptions are very important for two reasons: First is because it's acceptable by many application scenarios. For example, if an extractor system drives two triples from web page, then the correctness of these triples are independent. Second is because open world semantic is different than close world semantic which is used in almost all the previous works. Open world semantic allow to more independency work.

3.3.1 Quality of the Source

In this section we assume that the given sources are independent. Then we find the quality of the given sources. The quality of source is necessary because it affects the truthfulness of the triple. That means if the source has a high precision then the triple that it provides is more likely to be true. However, if the source has a high recall, then the triple that is not provided by the same source considers being false. Therefore, we start calculating the quality of the sources by using conditional probability, so that we can find the correctness of its information. Thus, to find the quality of the source, we need to define and find each source recall and precision.

Table 7. Summary of Notations Used in Fusing Data.

Notation	Description
S	Set of sources S= {S_1, … , S_n}
O_i	Set of output triples of source S_i
Ȯ	Ȯ ={O1, … , On}
Ȯt	Subset of observation in Ȯ that refer to triple t.
Pi (resp ps*)	Precision of source Si (resp. sources S*)
ri (resp rs*)	Recall of source Si (resp. sources S*)
qi (resp. qs*)	False positive rate of Si (resp. S*)
Si |= t	Si outputs t(t ∈ Oi)
S*|= t	∨ Si Є S*, Si |= t
Pr (t | Ȯ)	Correctness probability of triple t
Pr(t), Pr (¬t)	Pr(t= true) and Pr(t = false) respectively

Recall: Recall also called as positive predictive value, is the ratio of the number of relevant triples retrieved to the total number of relevant triples in the database.

r_i = pr (Si |= t | t) … (1)

Precision: Precision also called as sensitivity, is the ratio of the number of relevant triples retrieved to the total number of irrelevant and relevant triples retrieved. Figure 2 shows the standard meaning of precision and recall for a source.

$p_i = pr\ (t \mid Si \mid= t)$ … (2)

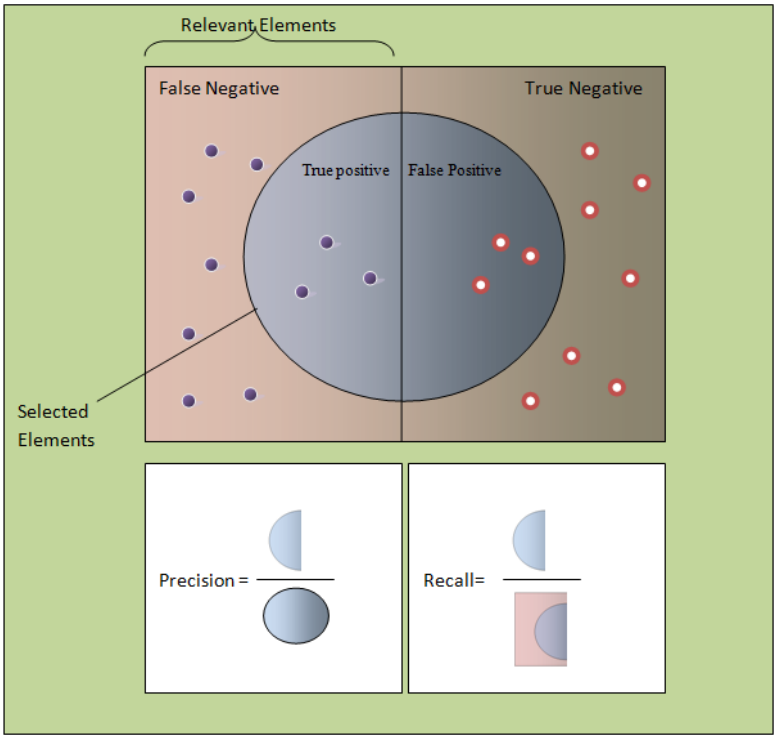

Figure 2. Precision and Recall

Example 1:- The precision of S2 is 3/7= 0.43 because we have just 3 out of 7 in O1 is correct. The recall of S2 is 3/6= 0.5 because O2 has just 3 out of 6 true triples. The precision and recall of all the independent sources is shown in table 8. Before start explaining the steps to find the probability of each triple, we need to describe some

important methods that lead to derive the equations such as Bayesian analysis and conditional probability.

Conditional Probability: conditional probability calculates the probability of occurring a given event A when other events occurred B. It is also called as a probability of A given B, which is denoted as: $P(A/B) = P(B \cap A) / P(B)$

Bayesian Method: also called Bayes' rule or Bayes' law, is a method of statistical Mathematic that deals with an event's probability depending on the event's condition. Bayes' rule denoted as:

$P(A \mid B) = (P(B \mid A). P(A))/ P(B)$, where $P(B \mid A)$ is the probability of B given A. whether P(A) and P(B) are probabilities of A and B respectively and independent of each other.

Table 8. Precision and Recall for Each Source.

Sources	Precision	Recall
S1	0.57	0.67
S2	0.43	0.5
S3	0.8	0.67
S4	0.67	0.67
S5	0.67	0.67

3.3.2 Estimate the Probability for Each Triple

Based on the quality of each source, we can compute the probability of each triple t to find if it's true or not. By using Bayes' rule to express Pr (t | O_t) based on the inverse probabilities Pr (Ot | t) and Pr (O_t | ¬t) which represents the probability of deriving the observed output data conditioned on t being true or false respectively. Also, the priori probability that triple t is true denoted as Pr (t) = α, where α initially assumes to be 0.5 in the given training dataset. Depending on the Bayes' Rule to derive pr (t| O_t):

$$\mathrm{pr}(t|O_t) = \frac{\mathrm{pr}\,(Ot \mid t).\,\mathrm{pr}(t)}{\mathrm{Pr}(Ot)} \ ,$$

and since $\mathrm{pr}(Ot) = (Ot \mid t).\,\mathrm{pr}(t) + \mathrm{pr}(Ot \mid \neg t).\,\mathrm{Pr}(\neg t)$, then

$$\mathrm{pr}(t|O_t) = \frac{\mathrm{pr}\,(Ot \mid t).\,\mathrm{pr}(t)}{(Ot \mid t).\,\mathrm{pr}(t) + \mathrm{pr}(Ot \mid \neg t).\,\mathrm{Pr}(\neg t)} \ , \text{and since } \mathrm{pr}(t) = \alpha$$

We will get the following equations:-

$$\mathrm{pr}(t|O_t) = \frac{\alpha \mathrm{Pr}\,(Ot \mid t)}{\alpha \mathrm{Pr}\,(Ot \mid t) + (1 - \alpha)\,\mathrm{Pr}\,(Ot \mid \neg t)} \qquad \ldots (3)$$

Moreover, the probabilities Pr (Ot | t) and Pr (Ot | ¬t) can be expressed using the true positive rate which called also as recall and the false positive rate which also called as complement of specificity, of each source as follows:

$$\mathrm{Pr}\,(Ot \mid t) = \prod_{Si \in St} \mathrm{Pr}\,(Si \mid = t \mid t) \prod_{Si \in St^-} \mathrm{Pr}\,(Si \mid = t \mid t) \qquad \ldots (4)$$

$$\Pr(Ot \mid t) = \prod_{Si \in St} \Pr(Si \mid\!= t \mid \neg t) \prod_{Si \in St^{-}} (1 - \Pr(Si \mid\!= t \mid \neg t)) \quad \ldots (5)$$

Where the set of sources that provide t denotes as St, and the set of sources that do not provide t denotes as St^{-}.

Now, we need to derive the false positive rate $q_i = \Pr(Si \mid\!= t \mid \neg t)$ for each source, which depends on the true positive rate and the precision for the specific source. The same as before, by using Bayes' rules on $\Pr(t \mid Si \mid\!= t)$ it gets the same eq. (3) and then by applying conditional probability for precision, recall, and false positive for each source, it gets:-

$$\text{pr}(t \mid Si \mid\!= t) \frac{\alpha \Pr(Si \mid\!= t \mid t)}{\alpha \Pr(Si \mid\!= t \mid t) + (1 - \alpha) \Pr(Si \mid\!= t \mid \neg t)}, \text{then}$$

$$\text{pi} = \frac{\alpha \text{ri}}{\alpha \text{ri} + (1 - \alpha)\, \text{qi}}, \text{then we get}$$

$$\text{qi} = \frac{\alpha}{1 - \alpha} . \frac{1 - \text{pi}}{\text{pi}} * \text{ri} \quad \ldots (6)$$

Calculating false positive rate is different than calculating the recall, because in false positive case it doesn't consider only the false triples that are providing by one source divided with the total number of false in the uncertain databases. Furthermore, one of the most important things to take into account in this case, is that false positive rate must be fall in the range of [0, 1] to be valid.

Example 2:- by taking α = 0.5, we can derive false positive rate for S2. Since its Precision= 0.47 and recall= 0.5 then q2= $\frac{0.5}{1-0.5} . \frac{1-0.47}{1.47} . 0.5 = 0.67$. Table 9 shows false positive rate for the five sources of the training data.

Table 9. False Positive Rate of the Five Sources.

Sources	False Positive Rate
S1	0.5
S2	0.67
S3	0.167
S4	0.33
S5	0.33

After finding the recall and false positive rate for each source, then we can find the correctness probability of output triple by following eq:

$$pr(t|O_t) = \frac{1}{1+\frac{1-\alpha}{\alpha}.\frac{1}{\mu}} \quad \text{... (7)}$$

Where μ can be computed based on the contribution of each source for each triple. Each source has a contribution $\frac{ri}{qi}$ for a triple that it provides and $\frac{1-ri}{1-qi}$ for the triple that it doesn't provide. Then for each triple in order to find the μ we multiply the contribution of the sources that provides it and don't provide it.

So, μ can be defined as the equation below:-

$$\mu = \prod_{Si \in St} \frac{ri}{qi} \prod_{Si \in St^{-}} \frac{1 - ri}{1 - qi} \quad \ldots (8)$$

Example 3:- to find the probability of triple 2 which is provided by S1 and S2 but not by other three sources, we apply equation (7). But we need to calculate the μ of the triple first:-

$$\mu = \frac{r1}{q1} \cdot \frac{r2}{q2} \cdot \frac{1 - r3}{1 - q3} \cdot \frac{1 - r4}{1 - q4} \cdot \frac{1 - r5}{1 - q5} = 0.1$$

Hence, the pr ($t2$, O_{t2}) = $\frac{1}{1 + \frac{1-0.5}{0.5} \cdot \frac{1}{0.1}}$ = 0.09, which means it's false.

However, independent sources can lead sometimes to wrong answers. For example, if we want to find the probability of triple 8, depending on the independent between sources, we get 0.62 since μ= 1.6 but triple 8 is in the fact is false. Thus, the correlation between sources could get more accurate results as we can see in the next section.

3.4 Fusing Correlated Sources

The correlation between sources affects the belief of triple truthfulness. Therefore, in this section we are keeping in mind the correlation between sources to find the probability of the given triple.

The results in this case significantly improve, if we compare it with the independent sources. The steps of finding the probability of the triples given correlated source are similar to the steps with the independent sources. Thus, we need to find the precision and the recall for correlated sources. Let denote ps* to the joint precision of sources S*, where S* is a group of sources that provides the given triple. ps* is represents the portion of triples in the output of all sources in S^* which are correct. Also, let denote rs* to the recall of sources which represents the portion of all correct triples that are output by all the sources in S^*. Hence, by following the equations below, we can find the joint precision and the joint recall for a group of sources:-

$$pS* = Pr\ (t \mid S* \mid= t) \quad \ldots (9)$$

$$rS* = Pr\ (S* \mid= t \mid t) \quad \ldots (10)$$

Where, the total number of joint precision and recall parameters for a set S of n sources can be calculated by taking the total of 2 (2^n-1).

Example 4: the subset of sources {S1, S4, and S5} provides t1, t6, t8, t9 and t10.

Therefore, their joint precision is $\frac{3}{5} = 0.6$, since they are providing three correct triples of total five common triples; and their joint recall is 3/6= 0.5 since they have three correct triples over total six correct triple in the whole training set. The table below shows the joint recall and joint precision for selected subset of sources.

Table 10. False Positive Rate for Selected Subset of Sources.

Correlated sources	Join precision	Joint recall
S2S3	0.67	0.33
S1S3	1	0.33
S1S2S4	0.33	0.167
S1S4S5	0.6	0.5

After finding the joint precision and joint recall for the subset of sources, we are able to find the probability of a given triple. There are two different approaches to find the probability by using correlated sources. They differ in their accuracy and accounting cost. The two approaches explained in the sections 3.4.1 and 3.4.2 which are the exact solution and aggressive solution.

3.4.1 Advantages of Correlation between Sources

There are a lot of advantages of correlation between sources more than increasing the belief of the truthfulness of the triple. As are mentioned below, the correlation helps in:-

- Not to be affected by coping between sources: for example, assume that all the source in a dataset coping from each other. Then, we can consider them as one source as their joint recall and joint false positive. In this case, $\mu_{corr} < \mu_{inde}$, which means false triple will get low probability with assuming correlation sources.

- Not to be affected by overlapping on true triples: for example, assume that all the sources in a dataset provides highly overlapping collection of true triples. However, each one of these sources has different mistakes. Then, by getting the joint recall as r_i and false positive as q^n, we get $\mu_{corr} > \mu_{inde.}$ This gives us more confidence in calculating the probability of true triples.
- Not to be affected by overlapping on false triples: for example, assume that the entire sources in a dataset provides highly overlapping collection of false triples. However, each one of these sources has different true triples. Then, by getting the joint recall as r^n and false positive as q, we get $\mu_{corr} < \mu_{inde.}$ This results in low probability for the false triples.
- Not to be affected by complementary source: for example, if the sources do not have high overlapping triples but their results are trustable, then in this case, by calculating joint recall r^n and false positive as q which is 0 we get that $\mu_{corr} \approx \infty$; Which means the true triples are highly trustable. This case also includes if one triple provided just by one source, then the correctness of the triple will not be penalized for that reason.

3.4.2 Positive Correlation and Negative Correlation

The correlation between two sources considers being positive correlation if the sources are nearly duplicates of each other. However, if the two sources have a less overlap between each other, then the correlation between them considers negative correlation. In the both situation, the triple that provided by one source or by both of them should not be affect by the correlation type.

3.4.3 Exact Solution

In this section we need to compute the equation (4) and (5) by considering the correlation between sources. The same as before, let us denote St to the set of sources that provide t and St^- to the set of sources that don't provide t. Then the equation became as follows:-

$$\Pr(O_t \mid t) = \Pr\left(\left(\bigwedge_{S \in St} S \mid = t\right) \wedge \left(\bigwedge_{S \in St^-} S \mid \neq t\right) \mid t \right) \quad \ldots (11)$$

Then by using inclusion-exclusion principle, the formula turns into:

$$\Pr(O_t \mid t) = \sum_{S* \subseteq St} (-1)^{|s*|} \Pr(\{ S_t \cup S^* \} \mid = t \mid t), \text{ then}$$

$$\Pr(O_t \mid t) = \sum_{S* \subseteq St} (-1)^{|s*|} r S_t \cup S^* \quad \ldots (12)$$

They are computing $\Pr(O_t \mid \neg t)$ in similar way to equation (12) but here by using joint false positive rate as follows:-

$$\Pr(O_t' \mid \neg t) = \sum_{S^* \subseteq S_{t^-}} (-1)^{|s^*|} qS_t \cup S^* \quad \ldots (13)$$

So we can compute $\Pr(O_t' \mid t)$, $\Pr(O_t' \mid \neg t)$ by Eq. (12) and (13). Now, we need to calculate the probability of the triple which is:

$$\Pr(t|O_t') = \frac{1}{1+\frac{1-\alpha}{\alpha}.\frac{1}{\mu}} \quad \ldots (14)$$

Where

$$\mu = \frac{\Pr(Ot' \mid t)}{\Pr(Ot' \mid \neg t)} \quad \ldots (15)$$

Example 5: triple t8 is provided by four sources S_{t8}= {S1, S2, S4, S5}.Thus, to compute $\Pr(O_t' \mid t)$ and $\Pr(O_t' \mid \neg t)$ depending on equations (12) and (13) respectively, we need to calculate joint recall for the sources that provides it r_{1245} and joint recall for all sources r_{12345}. Also, we need to calculate joint false positive for the sources that provides it q_{1245} and for all sources q_{12345}. Therefore, we get:-

$$\Pr(O_{t8}' \mid t8) = r_{1245} - r_{12345} = 0.167 - 0 = 0.167$$

$$\Pr(O_{t8}' \mid \neg t8) = q_{1245} - q_{12345} = 0.167 - 0 = 0.333$$

Thus, $\mu = \frac{0.333}{0.167}$ then we apply it to calculate $\text{pr}(O_t' \mid t_8) = \frac{1}{1+\frac{1-0.5}{0.5} \cdot \frac{0.333}{0.167}} = 0.334$

which is false. It's clear that by using correlation we can get more accurate results.

Exact solution is a useful and accurate approach to compute the probabilities of triples. However, it increases computational cost and especially if we have a large number of sources. Therefore, turning on aggressive approach is better because it's more practical approach to approximate Pr (Ȯt | t) and Pr (Ȯt | ¬t).

3.4.4 Aggressive Approximation

Aggressive approximation is a linear approximation that can reduce computational cost by enforcing a set of assumption. First, we find joint recall and joint false positive for a set of sources that provide each triple. Then we can compute the probability for each triple as:

$$\Pr(t|\bar{O}_t) = \frac{1}{1+\frac{1-\alpha}{\alpha}.\frac{1}{\mu_{aggr}}} \quad \ldots (16)$$

Where μ_{aggr} can calculate as follows:-

$$\mu_{aggr} = \prod_{Si \in St} \frac{C_i^+ ri}{C_i^- qi} \prod_{Si \in St^-} \frac{1 - C_i^+ ri}{1 - C_i^- qi} \quad \ldots (17)$$

Where

$$C_i^+ = \frac{r1 \ldots n}{ri.r12 \ldots (i-1)(i+1) \ldots n} \quad \ldots (18)$$

$$C_i^- = \frac{q1 \ldots n}{qi.q12 \ldots (i-1)(i+1) \ldots n} \quad \ldots (19)$$

Where, aggressive approximation just uses (2n+1) instead of $2(2^n - n - 1)$ correlation parameters.

Note that in case of independent sources $C_i^+ = C_i^- = 1$. In this case the approximation gets the same result as eq (7).

Example 6: triple t8 is provided by S_{t8}= {S1, S2, S4, S5} and not provided by source S3. Then, we can compute the probability of triple t8 by aggressive approximation as illustrated in the steps below:-

First we compute μ_{aggr} :

$$\mu_{aggr} = \frac{0.67 \cdot 0.5 \cdot (1 - 0.75 \cdot 0.67) \cdot 1.5 \cdot 0.67 \cdot 1.5 \cdot 0.67}{2 \cdot 0.5 \cdot 0.67 \cdot (1 - 0.167) \cdot 3 \cdot 0.33 \cdot 3 \cdot 0.33} = 0.3$$

Then the probability of triple t8 is pr ($t_8 | \dot{O} t_8$) = $\frac{1}{1 + \frac{0.5}{1-0.5} \cdot \frac{1}{0.3}} = 0.23$, which means t8 is false. Thus, aggressive approach also can get the correct results and more accurate than the exact solution. Furthermore, the computational in this case is linear in the number of sources.

3.4.5 The Relation between Sources and Triples

The quality of a source depends on the correctness of the triples that it provides. Where, the source consider to be a good source, if it is more likely provides a true triple than false triples, Pr (Si |= t | t) > Pr (Si |= t | ¬t). Also, this affects the correctness of the

triple such that if the source is a good source then more likely the triple that it provides is true otherwise it is false. To illustrate more, Let S'= S ∪ {S'} and O' = Ȯ ∪ {O'} then:

1. If S' is a good source:
    - If S' |= t, then Pr (t | Ȯ' t) > Pr (t | Ȯ t).
    - If S' |≠ t, then Pr (t | Ȯ' t) < Pr (t | Ȯ t).
2. If S 0 is a bad source:
    - – If S 0 |= t, then Pr (t | Ȯ' t) < Pr (t | Ȯ t).
    - – If S 0 |≠ t, then Pr (t | Ȯ' t) > Pr (t | Ȯ t).

CHAPTER IV

PROBABILITIES FOR UNCERTAIN DATA FUSING

In this chapter, we have identified the most important ideas that helped us to successfully calculate the probabilities of uncertain data by using fusing algorithms. We introduce our idea of fusing uncertain data with probabilities that will form the basis for computing the probabilities of the triples in the given databases. Also, we introduce some new notations in addition to the notations in table 7.

4.1 Calculating Probabilities in Data Fusion

Given n number of sources, let us assume that all of these sources provide triples with assigned probabilities. Now, we present two methods to calculate triples probabilities, one by considering independencies between the sources and then by considering the correlation between them.

4.1.1 Considering the Independencies between the Sources

- We initialize by calculating the initial probability of the triples α_1. For each triple, we compute the probabilities that are given by each source and we used three different methods for that:-

1. An ad-hoc approach that uses the average of confidence
2. The fuzzy logic approach that uses the maximum of the confidence
3. The probabilistic theory approach that computes the prior probability of each fact assuming sources are independent.

- The third approach worked best in our experiments. To illustrate the third approach, suppose we have 5 sources and two of them are providing a triple with 0.2 and 0.3 probabilities respectively, then the initial probability is 1- (1-0.2) * (1- 0.3) = 0.25 and we consider it as false.
- We choose the value of correctness threshold $\hat{p}$ which is the key factor in the probabilistic data fusion. When probability of the data is less than $\hat{p}$ is considered being false and those with probabilities equal or higher than $\hat{p}$ are considered to be true.

Note that, the choice of $\hat{p}$ affects the performance of data fusion greatly. A correctness threshold that is too high can filter out the majority of facts that are not true (hence a low false-positive rate for the algorithm - which is desirable), but can also miss the majority of true facts (hence a low recall rate for the algorithm- which is undesirable). On the other hand, a correctness threshold that is too low can find the majority of true facts (hence a high recall rate for the algorithm - which is desirable), but also render the majority of false facts to be true (hence a high false-positive rate for the algorithm- which is undesirable). Furthermore, many data mining systems dismiss mined facts with a confidence lower than a fixed number, such as 90% (for example, see [24]).

- After getting the initial probability for each triple $\alpha_{1,}$ and $\hat{p}$, we start to find source quality by calculating its precision and recall by using equation (1) (2) from chapter III.
- We calculate the false positive but in this case by depending on the common method of calculating it, which is by considering the number of false in each source to the total number of false that is in the database as the equation below:

$$q_i = pr\ (Si \models \neg t \mid \neg t) \quad \ldots (20)$$

As an example, assume we have a source S that provides 5 false triples and there are 10 false triples in the dataset. Then, $q_s = \frac{5}{10} = 0.2$.

- Then we start to estimate each triple's probabilities by using equation (7) but before that we need to find the value of μ but in this case by using α_1 instead of α.

4.1.2 Considering the Correlation between the Sources

We are using two methods to compute the probability of the triples which are the exact solution and the aggressive solution. The reason of using two methods is as we mentioned in chapter III, is that the exact solution is sometimes with large number of sources is not practical when we have a large number of source since it increases the cost of computational [5]. It's working much stronger with small number of sources. Therefore, we present aggressive approximation method which is working excellently with different number of sources. The steps of the both method is almost similar to the

independent sources one except for the way that we are considering here the correlation between the sources that provide each triple.

4.1.2.1 Exact Solution

- We initialize by calculating the initial probability of the triples α_1. For each triple, we use the probabilistic theory approach that computes the prior probability of each fact assuming sources are independent.
- We choose the value of correctness threshold $\hat{p}$ which is the key factor in the probabilistic data fusion. When the probability of the data is less than $\hat{p}$ are considered being false and those with probabilities equal or higher than $\hat{p}$ are considered to be true.
- After getting the initial probabilities for each source, we start to find the correlated source quality by calculating their precision and recall as equation (9) (10).
- We calculate false positive of the correlated sources. In this case, also we depend on the common method of calculating the false Positive, which is by considering the number of false in each group of correlated source to the total number of false that is in the database as the equation below:

$$q_{s*} = pr\ (S^* \models \neg t \mid \neg t) \quad \ldots (21)$$

- We compute Pr (Ȯt |t), Pr (Ȯt |¬t) by using equations (12) and (13) respectively.

- Then we start to estimate each triple's probabilities by using equation (14) but before that we need to find the value of μ but in this case by using α1 instead of α in equation (15).

4.1.2.2 Aggressive Approximation

- We initialize by calculating the initial probability of the triples α_1. For each triple, we use the probabilistic theory approach that computes the prior probability of each fact assuming sources are independent.
- We choose the value of correctness threshold $\hat{p}$ which is the key factor in the probabilistic data fusion. When the probability of the data is less than $\hat{p}$ are considered being false and those with probabilities equal or higher than $\hat{p}$ are considered to be true.
- After getting the initial probabilities for each source, we start to find the correlated source quality by calculating their precision and recall as equation (9) (10).
- We calculate false positive of the correlated sources. In this case, also we depend on the common method of calculating the false Positive, which is by considering the number of false in each group of correlated source to the total number of false that is in the database as equation (21).
- We compute C_i^+, C_i^- for each source that provides the triple by using equations (18) and (19) respectively.

- Then we start to estimate each triple's probabilities by using equation (16) but before that we need to find the value of μagg but in this case by using α1 instead of α as equation (15).
- The result that we get from aggressive approximation is very accurate especially when number of sources is very big. So, with a large number of sources, it can be more powerful even more than exact solution technique.

4.2 Example for Counting the Probabilities

We present example to elucidate the calculation of probabilities by using the two methods that which explained in the previous sections. First, we calculate the probabilities by considering independent between sources. Then, we consider the correlation between the given sources. Consider 5 Sources where each one of them provides 5 triples which assigned to random probabilities as it's shown in table 11.

Table 11. Mini Dataset with 5 Sources and 9 Triples.

ID	Country	Capitol	Correct	S1	S2	S3	S4	S5
1	United Arab	Abu Dhabi	0.960	0.37	0.34	0	0.05	0.9
2	Nigeria	Abuja	0.81	0.12	0.57	0.15	0.42	0
3	Ghana	Accra	0.99	0.2	0	0.11	0.03	0.9
4	Ethiopia	Addis	0.99	0.89	0	0	0.85	0.6
5	Algeria	Algiers	1	0.09	0.65	0.71	0.71	1
6	United Arab	f2	0.69	0	0	0.69	0	0
7	Nigeria	f1	0.77	0	0	0	0	0.7
8	Ghana	f3	0.22	0	0.23	0	0	0
9	Ethiopia	f3	0.88	0	0.73	0.58	0	0

4.2.1 Considering the Independencies between Given Sources

In the beginning the algorithm calculates the initial probabilities of each triple which identify as α_1. For example, triple t1 provided by 4 sources S1, S2, S4, and S5 and not provided by source S3, then the initial probability is 1-(1-0.37)*(1-0.34)* (1-0.05) *(1-0.9) = 0.960499 and we consider it is true if its greater or equal to $\hat{p}$ else its false. After that the algorithm starts to find source quality by calculating its precision and recall by using equation (1) (2) as it is shown in table 12.

Table 12. Precision, Recall, and False Positive Rate for the 5 Sources.

Sources	Precision	Recall	False Positive Rate
S1	0.8	1	0.2
S2	0.4	0.5	0.6
S3	0.4	0.5	0.6
S4	0.8	1	0.2
S5	0.8	1	0.2

After that, it calculates the false positive for each source by using equation (20). To illustrate more, the false positive for source S1 is $\frac{1}{4}$ = 0.2. Now, we begin to estimate each triple's probabilities by using equation (7) and before that we need to find the value of μ. For example, for triple t8, μ =0, pr (t_8 |Ȯt_8) = $\frac{1}{1+\frac{0.2299999}{1-0.2299999}\cdot\frac{1}{0}}$ = 0 which is consider as false. We did one experiments by taking $\hat{p}$ = 0.95 and the result we get is

- The Experiment Recall is 0.8
- The Experiment False Positive is 0.0
- The Experiment Precision is 1.0

4.2.2 Considering the Correlation between the Sources

4.2.2.1 Exact Solution

The algorithm initialize by calculating the initial probability of the triples α_1. Then, it starts to find the correlated source quality by calculating their precision and recall as equation (9) and (10) as some of them shown in table 13. Also, for the correlated sources we need to calculate false positive rate by using equation (21).

Table 13. Recall and False Positive Rate for the Selected Subset of Sources.

Sources	Recall	False Positive
S1S3	0.5	0.2
S1S4S5	1	0
S1S2S4S5	0.5	0
S2S3S4S5	0.25	0
S1S2S3S4S5	0.25	0

We compute Pr (Ȯt | t), Pr (Ȯt | ¬t) by using equations (12) and (13) respectively. For example, triple t1 has Pr (Ȯt | t1) =0.25, Pr (Ȯt | ¬t1) = 0.

Then we estimate each triple's probabilities by using equation (14) and before that we need to find the value of μ by using equation (15). For example, for triple t7 we first find the initial probability which is 0.77. Then, we find Pr (Ȯt | t7) = 0.5 and Pr (Ȯt|¬t7) =

0.2. So, $\mu = \frac{0.5}{0.2} = 0.25$, and pr $(t_9 | \dot{O}t_9) = \frac{1}{1+\frac{1-0.77}{1-0.77} \cdot \frac{1}{0.25}} = 0.893$ which considers false.

We did one experiments by taking $\hat{p} = 0.95$ and the result we get:-

- The Experiment Recall is 0.8
- The Experiment False Positive is 0.0
- The Experiment Precision is 1.0

4.2.2.2 Aggressive Approximation

We initialize by calculating the initial probability of the triples α_1. Then, we start to find the correlated source quality by calculating their precision, recall, and false positive of the correlated sources as is done in exact solution.

We compute C_i^+, C_i^- for each set of source by using equations (18) and (19) respectively. For example, triple t1 is providing by S1, S2, S4, and S5 but not by source S3. So, we have to find C_i^+ and C_i^- for all the sources. As an illustration, for source S1 $C_1^+ = \frac{rS1S2S3S4S5}{rS1 * r\ S2S3S4S5}$ and $C_1^- = \frac{qS1S2S3S4S5}{qS1 * q\ S2S3S4S5}$. After that, we start to estimate each triple's probabilities by using equation (16) but before that we need to find the value of μ_{agg} by using equation (15). We did one experiments by taking $\hat{p} = 0.95$ and the result we get:-

- The Experiment Recall is 1.0
- The Experiment False Positive is 0.0
- The Experiment Precision is 1.0

CHAPTER V

DATA SETS

5.1 Training Data Set

The training dataset that we used is taken from [5] in which is taking from Barak Obama's Wikipedia page using five different extractor systems as it's shown in figure 3. The data is in the form of (subject, predicate, object) such as {Obama, spouse, Michelle}. The dataset consist of 10 triples. Each triple has its correctness value (Yes/No), and the sources that provides each triple as it's shown in table 14. Where, the check marks mean that the source are providing the knowledge of that triple. For instance, triple 2 is provided by the two sources S1 and S2 but not by other sources.

Figure 3. The Wikipedia Page of Barak Obama and Five Extractors that Extract Knowledge from it.

Table 14. Data Extracted by Five Different Extractors from the Wikipedia Page for Barack Obama.

ID	Knowledge Triple	Correct	S1	S2	S3	S4	S5
T1	{Obama, Profession, President}	Yes	√	√		√	√
T2	{Obama, died, 1982}	No	√	√			
T3	{Obama, Profession, lawyer}	Yes			√		
T4	{Obama, religion, Christian}	Yes		√	√	√	√
T5	{Obama, age,50}	No		√	√		
T6	{Obama, support, White Sox}	Yes	√			√	√
T7	{Obama, Spouse, Michelle}	Yes	√	√	√		
T8	{Obama, administered by, John}	No	√	√		√	√
T9	{Obama, operation, 2011}	No	√	√		√	√
T10	{Obama, Profession,c.organizer}	Yes	√		√	√	√

5.2 Countries and Capitals Data Set (Intentional False)

We created an algorithm to produce two-dimensional data set. The data is in the form of (country name, capitol name) such as {Iraq, Baghdad}. The data set consists of 201 correct triples and a certain number of false triples depending on the number of false value that the user inserts. Each triple has its correctness value, which is the initial probability that calculated with the given probabilities by each source as it's shown in the table 15.

The steps of the algorithm to create a data set (with n number of sources which can be decided by the user) are as follows:

1. We determine the false rate which it can be between 0.1 and 0.9.
2. Then we determine the false value, how many false capitols we want to have in the dataset.
3. For each one of the given sources, the algorithm gives random numbers between 0.1 and 0.9 for all of the triples.
    - If the range is less than false rate, it puts 0 and inserts new tripe with the same country name but with different capitol name which is chosen from false value intentionally for all the sources who provide probability less than false rate; and then it gives random probability for that source. So, this source considers as a provider of the wrong triple.
    - If the range is greater or equal to the false rate, it gives a random probability .So, this source considers as provider to the correct triple.
4. Then for each triple, it calculates the average of the probabilities which is provided by the sources.

For example, if we want to create a data set with 5 sources, and we decide the false rate to be 0.2 and the false value as f1, f2 and f3. The algorithm gives random number for each source between 0.1 and 0.9 for all triples. Since the false rate is 0.2, so if the range of triple in each source is greater than 0.2, it gives a random probability. However, if the

range is less than 0.2, it puts a 0 and then adds new triple with the same country name but with different capitol and it will add random probability to them.

So, each source provides all the countries but not all the triples. For example triple 1 is provided by the sources S1, S2 and S5 but not by other sources. Creating different dataset by using this algorithm helps in validating the work of fusing methods.

Table 15. The First 8 Triples of Countries and Capitals (Intentional False) Dataset.

ID	Country	Capitol	Correct	S1	S2	S3	S4	S5
1	United A	Abu D	0.9700	0.76	0.11	0	0	0.86
2	Nigeria	Abuja	0.9998	0.99	0.07	0.75	0.77	0.67
3	Ghana	Accra	0.941	0.7	0	0.77	0.15	0
4	Ethiopia	Addis A	0.8600	0.02	0.36	0.23	0.37	0.54
5	Algeria	Algiers	0.9888	0.6	0.95	0.04	0.42	0
6	United A	f1	0.8905	0	0	0.85	0.27	0
7	Ghana	f2	0.831	0	0.35	0	0	0.74
8	Algeria	f1	0.71	0	0	0	0	0.71

5.3 Countries and Capitals Data Set (Random False)

We created another algorithm to produce two-dimensional data set but in this case with using random false. That means, there is no copying between the sources. The data is in the form of (country name, capitol name) such as {Iraq, Baghdad}. The data set consists of 201 correct triples and a certain number of false triples depending on the

number of false values that is given by the user. Each triple has its correctness value, which is calculated with the given probabilities by each source as it's shown in table 16. The algorithm to create dataset (with n number of sources which can be decided by the user) is:

1. We have to determine the false rate which it can be between 0.1 and 0.9.
2. Then, we have to determine the false value, how many false capitals we want to have in the dataset.
3. For each one of the given sources, the algorithm gives random numbers between 0.1 and 0.9 for all of the triples.
    - If the range is less than false rate, it puts 0 and inserts new tripe with the same country name but with different capitol name which is chosen from the given false value; and then it gives random probability for that source. So, this source considers as a provider to the wrong triple.
    - If the range is greater or equal to the false rate, it gives a random probability. So, this source considers as provider to the correct triple.
4. Then for each triple, it calculates the average of the probabilities which is provided by the sources.

For example, if we want to create a data set with 5 sources, and we decide the false rate to be 0.2. The algorithm gives random number for each source between 0.1 and 0.9 for all triples. Since the false rate is 0.2, so if the range of triple in each source is

greater than 0.2, it gives a random probability. However, if the range is less than 0.2, it puts a 0 and then adds new triple with the same country name but with different capitol. So, each source provides all the countries but not all the triples. For example triple 1 t1 is provided by the sources S1, S2, S4 and S5 but not by other sources. Creating different dataset by using this algorithm helps in validating the work of fusing methods.

Table 16. The First 9 Triples of Countries and Capitals (Random False) Dataset.

ID	Country	Capitol	Correct	S1	S2	S3	S4	S5
1	United Arab	Abu Dhabi	0.960499	0.37	0.34	0	0.05	0.9
2	Nigeria	Abuja	0.8134488	0.12	0.57	0.15	0.42	0
3	Ghana	Accra	0.9930936	0.2	0	0.11	0.03	0.99
4	Ethiopia	Addis	0.993895	0.89	0	0	0.85	0.63
5	Algeria	Algiers	1	0.09	0.65	0.71	0.71	1
6	United Arab	f2	0.69	0	0	0.69	0	0
7	Nigeria	f1	0.77	0	0	0	0	0.77
8	Ghana	f3	0.229999	0	0.23	0	0	0
9	Ethiopia	f3	0.886599	0	0.73	0.58	0	0

5.4 Toy Example Data Set

We created another example data by assuming probabilities for the training dataset that is given in [5] which we described in section 5.1. The data is in the form of (subject, predicate, object) such as {Obama, spouse, Michelle}. The data set consists of 10 triples. Each source provides some triples with a certain probability. The dataset contains five sources. So if the source provides a triple, we denote it as its given probability or as 0. The correctness value of the triples in the dataset calculated with the given probabilities by each source as it's shown in table 17.

Table 17. Data Extracted by Five Different Extractors from the Wikipedia Page for Barack Obama with Probabilities Added to it.

ID	Knowledge Triple	Correct	S1	S2	S3	S4	S5
T1	{Obama, Profession, President}	0.95192	0.66	0.43	0.06	0.71	0.09
T2	{Obama, died, 1982}	0.99512	0.89	0.23	0.55	0.82	0.29
T3	{Obama, Profession, lawyer}	0.99343	0.54	0.39	0	0.74	0.91
T4	{Obama, religion, Christian}	0.99451	0.62	0.97	0.42	0	0.17
T5	{Obama, age,50}	0.9127	0.4	0.68	0.23	0.41	0
T6	{Obama, support, White Sox}	0.9949	0.27	0.45	0.93	0.8	0.1
T7	{Obama, Spouse, Michelle}	0.9912	0	0.64	0.9	0.46	0.55
T8	{Obama, administered by, John G.}	0.99556	0.98	0.4	0.12	0.58	0
T9	{Obama, Surgical operation, 05/01/2011}	0.94639	0.54	0.02	0.59	0	0.71
T10	{Obama, Profession, community organizer}	0.72613	0.19	0	0	0.51	0.31

CHAPTER VI

RESULTS AND FINDINGS

In this work, fusing data techniques that we used showed significantly results by using several datasets. These results have been compared with each other to observe which technique performs better in term of accurate results and CPU time. Also, the method that we presented to calculate the probabilities of the triples using the probability that is given by the sources which provides certain number of triples, made the techniques more general; since it doesn't need for training dataset anymore. Thus, it can work with any type of datasets and with any number of sources and triples. In this chapter, we compared the probability of triples that result by considering independency and the correlation between the given sources to see which one works better. Also, we show the difference in result when we use different methods to calculate the initial probability for each triple.

6.1 Results

Our verification methodology includes comparing probabilistic of triples by considering independency and correlation between the sources.

First of all, we initialize by fusing uncertain data by taking into account the independency between the sources. We compute the probability of each triple after calculating its initial probability from sources that provides the triple.

Second of all, we fuse uncertain data by bearing in mind the correlation between the sources. Then, we compute the probability of each triple after calculating its initial probability from the sources that provides the triple. In correlation case, we have two methods which are exact solution and aggressive approximation. We use both of the methods to get more accurate results.

Finally, we compare the result that we get in each time to see which technique works better.

6.2 Independence Case

We did different experiments by considering independency between the sources. **First**, we show the different experiments by using Countries and capitals Dataset (Random False) as it's described in section 5.3. We used different number of source and calculate the precision, recall, and the false positive in each experiment in order to figure out the quality of the technique as its shown below:-

1- By using dataset with false rate = 0.1, which means number of false triples is just 10% in the whole data set. We used different number of sources and correctness value to understand the efficiency of the techniques.

   By using 10, 20, 30, 40, and 50 sources and correctness value, threshold, as almost 0.99, we get the results that are shown in the table 18. To make the results more clearly, we provide the diagrams for the precision, recall, false positive rate, execution time, and the number of sources in figure 4.

Figure 4. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 10% Random Errors.

Table 18. The Average CPU Time of Five Times Running and Sources p, r, and q by Considering Independency between the Sources Using Different Number of Sources.

Number of Sources	Number of Triples	CPU time	I/o time	Recall	False Positive	Precision
10	562	1,260	718	0.995	0	1
20	742	1,973	1,293	1	0.009225	0.9756097
30	938	3,011	2,187	1	0.005420	0.9803921
40	1004	4,592	4,156	1	0.004975	0.9803921
50	1070	6,073	4,952	1	0.013793	0.9433962

2- By using dataset with false rate = 0.2, which means number of false triples is just 20% of the whole data set. We used different number of sources and correctness value to understand the efficiency of the techniques.

By using 10, 20, 30, 40, and 50 sources and correctness value as 0.95, we get the results that are shown in the table 19. Also, we provide the diagrams for the precision, recall, false positive ate, execution time, and the number of sources in figure 5 to make the result more clear.

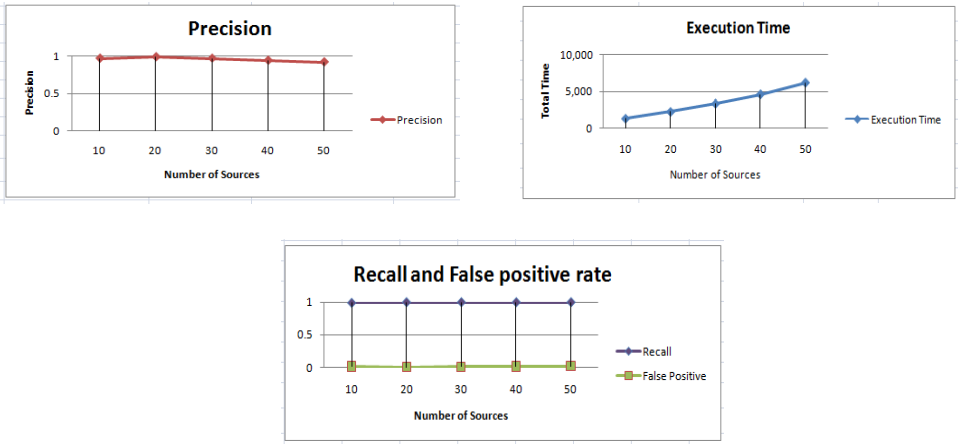

Figure 5. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 20% Random Errors.

Table 19. The Average CPU Time of Five Times Running and Sources p, r, and q for Different Number of Sources.

Number of Sources	Number of Triples	CPU time	I/o time	Recall	False Positive	Precision
10	663	1,278	622	0.99	0.012958	0.970588
20	927	2,231	1,570	1	0.002751	0.99009
30	1037	3,352	2,531	1	0.007168	0.970873
40	1120	4,604	3,702	1	0.013043	0.943396
50	1150	6,185	5,250	1	0.017894	0.921658

3- By using dataset with false rate = 0.3, which means number of false triples is just 30% of the whole data set. By using 10, 20, 30, 40, and 50 sources and correctness value as 0.95, we get the results that are shown in the table 20.

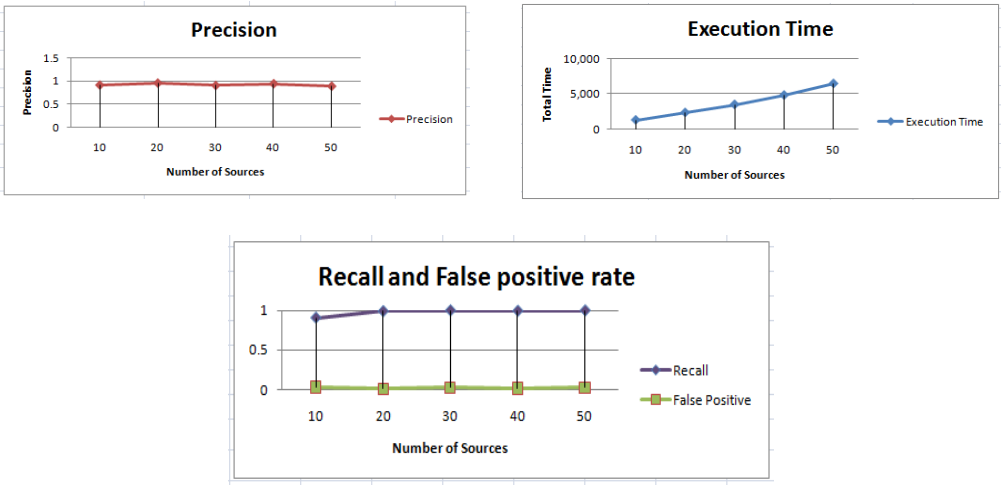

Figure 6. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 30% Random Errors.

Table 20. The Average CPU Time of Five Times Running and Sources p, r, and q for Different Number of Sources.

Number of Sources	Number of Triples	CPU time	I/o time	Recall	False Positive	Precision
10	770	1,253	640	0.905	0.029824	0.914141
20	1022	2,317	1,453	0.99	0.010948	0.956521
30	1120	3,466	2,906	1	0.022826	0.904977
40	1158	4,806	3,834	0.995	0.012526	0.943127
50	1184	6,463	5,434	1	0.025406	0.888888

4- By using dataset with false rate = 0.4, which means number of false triples is just 40% of the whole data set. By using 10, 20, 30, 40, and 50 sources and correctness value as 0.95, we get the results that are shown in the table 21.

Figure 7. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 40% Random Errors.

Table 21. The Average CPU Time of Five Times Running and Sources p, r, and q for Different Number of Sources.

Number of Sources	Number of Triples	CPU time	I/o time	Recall	False Positive	Precision
10	830	1,496	863	0.86	0.074603	0.785388
20	1073	2,301	1,499	0.96	0.027491409	0.888888
30	1146	3,553	2,718	0.99	0.022198	0.904109
40	1191	4,716	3,859	1	0.027245	0.881057
50	1194	6,322	5,343	0.995	0.035211	0.850427

5- By using dataset with false rate = 0.5, which means number of false triples is 50% of the whole data set. By using 10, 20, 30, 40, and 50 sources and correctness value as 0.95, we get the results that are shown in the table 22.

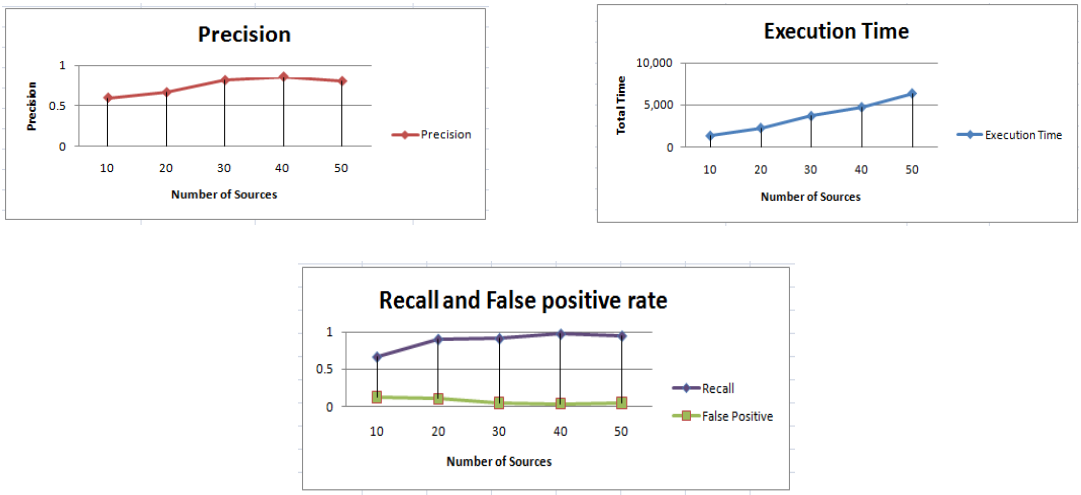

Figure 8. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 50% Random Errors.

Table 22. The Average CPU Time of Five Times Running and Sources p, r, and q for Different Number of Sources.

Number of Sources	Number of Triples	CPU time	I/o time	Recall	False Positive	Precision
10	728	1,352	749	0.665	0.123626	0.59641
20	1120	2,289	1,499	0.905	0.1	0.663003
30	1181	3,683	2,781	0.915	0.042813	0.813333
40	1196	4,734	3,843	0.975	0.03313	0.855263
50	1196	6,358	5,206	0.95	0.04718	0.80168

We calculate the CPU time in each case to show the effectives of the technique. The CPU time includes time it takes to read the data from the excel file, to print the results, and to do the operations to find the probabilities of each triple.

Second, we show the different experiments by using Countries and capitals Dataset (Intentional False) as it's described in section 5.2. We used different number of source and calculate the precision, recall, and the false positive in each experiment in order to figure out the quality of the technique as its shown below:-

1- By using dataset with false rate = 0.1, which means number of false triples is just 10% in the whole data set. We used different number of sources and correctness value to understand the efficiency of the techniques.

   By using 10, 20, 30, 40, and 50 sources and correctness value as almost 0.99, we get the results that are shown in the table 23.

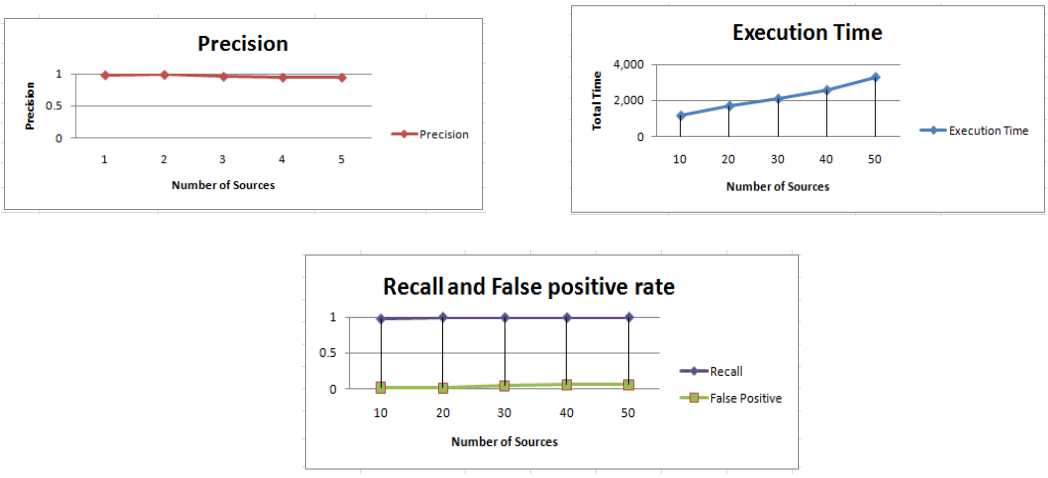

Figure 9. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 10% Intentional Errors.

Table 23. The Average CPU Time of Five Times Running and Sources p, r, and q for Different Number of Sources.

Number of Sources	Number of Triples	CPU Time	I/o Time	Recall	False Positive	Precision
10	385	1,176	562	0.975	0.01621	0.98484
20	395	1,718	1,015	1	0.0102	0.9900
30	400	2,107	1,390	0.995	0.04	0.96135
40	400	2,586	1,890	0.995	0.055	0.94761
50	400	3,301	2,577	1	0.055	0.94786

2- By using dataset with false rate = 0.2, which means number of false triples is just 20% of the whole data set. By using 10, 20, 30, 40, and 50 sources and correctness value as 0.95, we get the results that are shown in the table 24.

Figure 10. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 20% Intentional Errors.

Table 24. The Average CPU Time of Five Times Running and Sources p, r, and q for Different Number of Sources.

Number of Sources	Number of Triples	CPU time	I/o time	Recall	False Positive	Precision
10	395	1,145	593	0.85	0.18461	0.82524
20	400	1,562	953	0.96	0.065	0.93658
30	400	2,103	1,562	0.96	0.06	0.94117
40	400	2,598	1,859	0.995	0.025	0.9754
50	400	3,166	2,390	0.985	0.09	0.9162

3- By using dataset with false rate = 0.3, which means number of false triples is just 30% of the whole data set. By using 10, 20, 30, 40, and 50 sources and correctness value as 0.95, we get the results that are shown in the table 25.

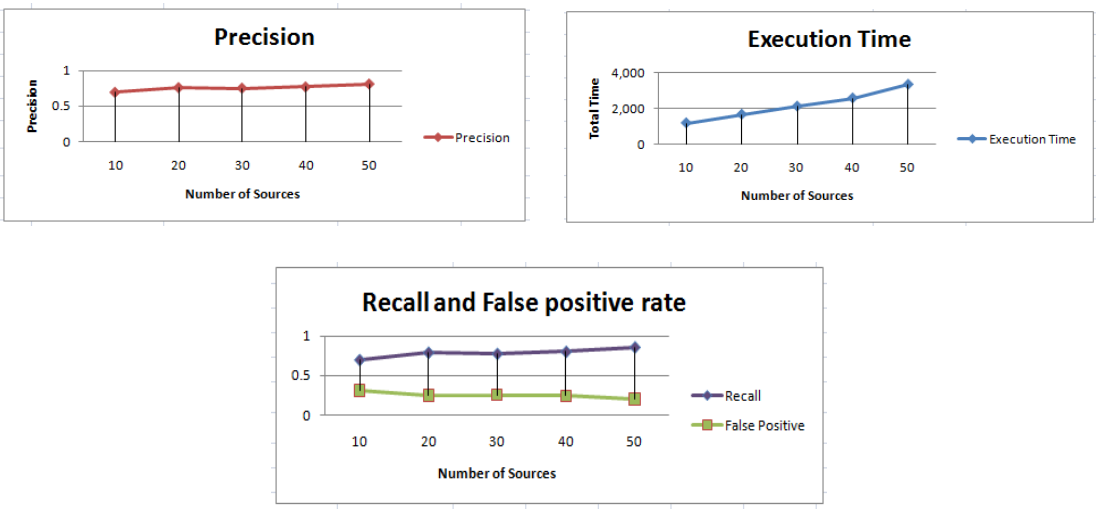

Figure 11. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 30% Intentional Errors.

Table 25. The Average CPU Time of Five Times Running and Sources p, r, and q for Different Number of Sources.

Number of Sources	Number of Triples	CPU time	I/o time	Recall	False Positive	Precision
10	399	1,171	578	0.7	0.311	0.6930
20	400	1,651	968	0.79	0.25	0.7596
30	400	2,104	1,453	0.775	0.255	0.7524
40	400	2,557	1,890	0.8	0.24	0.7692
50	400	3,317	2,624	0.855	0.2	0.8104

4- By using dataset with false rate = 0.4, which means number of false triples is just 40% of the whole data set. By using 10, 20, 30, 40, and 50 sources and correctness value as 0.95, we get the results that are shown in the table 26.

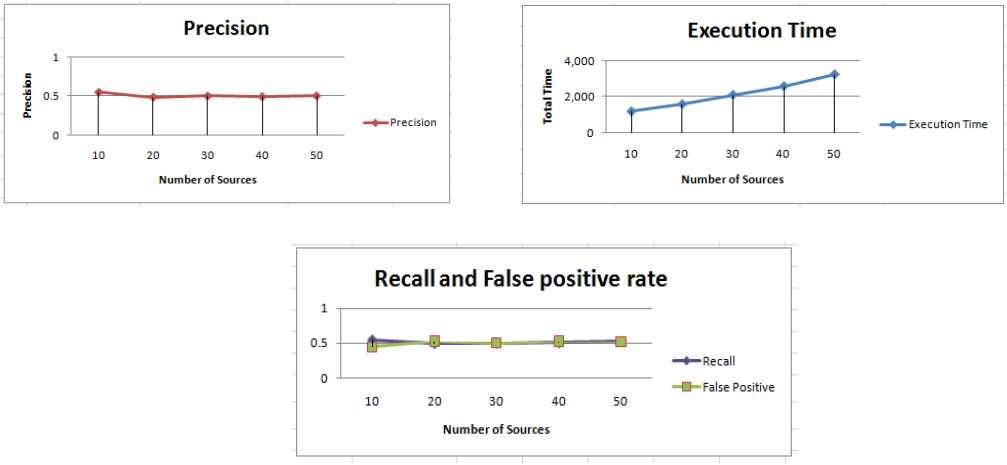

Figure 12. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 40% Intentional Errors.

Table 26. The Average CPU Time of Five Times Running and Sources p, r, and q for Different Number of Sources.

Number of Sources	Number of Triples	CPU time	I/o time	Recall	False Positive	Precision
10	400	1,182	562	0.545	0.445	0.5505
20	400	1,583	937	0.495	0.525	0.48529
30	400	2,099	1,389	0.5	0.5	0.5
40	400	2,580	1,937	0.505	0.525	0.49029
50	400	3,234	2,541	0.525	0.52	0.50239

5- By using dataset with false rate = 0.5, which means number of false triples is just 50% of the whole data set. By using 10, 20, 30, 40, and 50 sources and correctness value as 0.95, we get the results that are shown in the table 27.

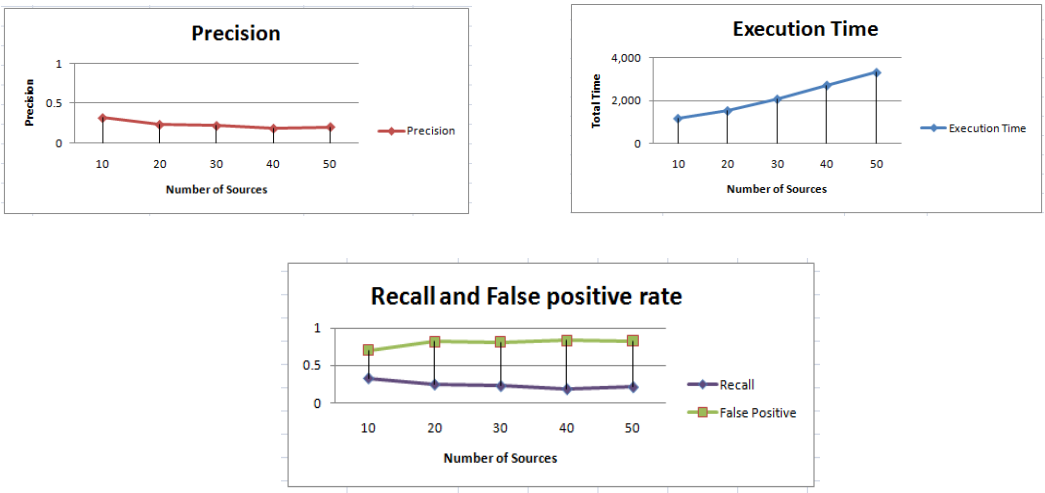

Figure 13. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 50% Intentional Errors.

Table 27. The Average CPU Time of Five Times Running and Sources p, r, and q for Different Number of Sources.

Number of Sources	Number of Triples	CPU time	I/o time	Recall	False Positive	Precision
10	400	1,160	609	0.325	0.705	0.31553
20	400	1,520	930	0.245	0.82	0.23004
30	400	2,067	1,343	0.225	0.815	0.21634
40	400	2,718	1,999	0.185	0.835	0.18137
50	400	3,317	2,530	0.21	0.83	0.20192

6.3 Correlation Case

As it mentioned before, there are two methods in the correlation case which are exact solution and aggressive approximation. We present each of them with the result of the experiments that done using both of the techniques.

6.3.1 Exact Solution

We did different experiments by considering correlation between the sources using exact solution method.

First, we show the different experiments by using Countries and capitals Dataset (Random False) as it's described in section 5.3. We used different number of source and calculate the precision, recall, and the false positive in each experiment in order to figure out the quality of the technique as its shown below:-

1- By using dataset with false rate = 0.1, which means number of false triples is just 10% of the whole data set. We used different number of sources and correctness value to understand the efficiency of the techniques. By using 10, 20, and 30 sources and correctness value as almost 0.99, we get the results that are shown in the table 28.

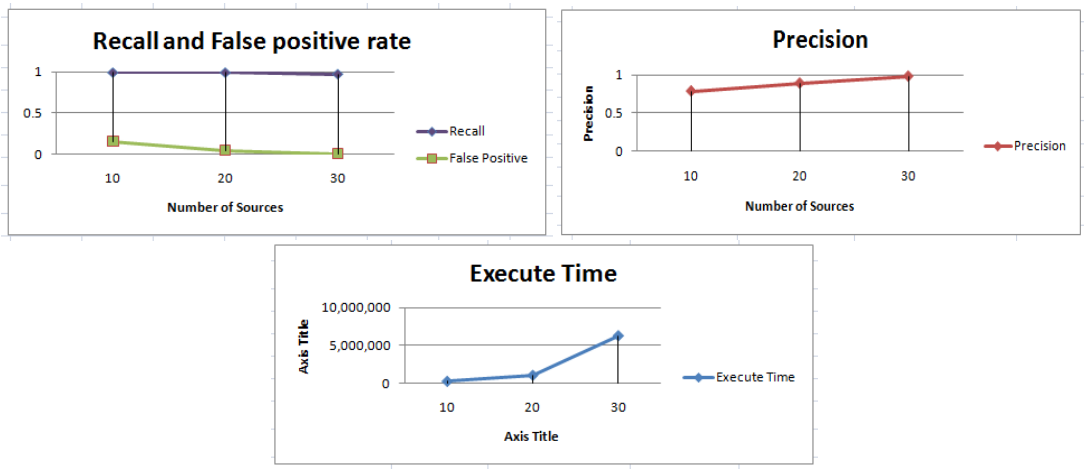

Figure 14. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 10% Random Errors.

Table 28. CPU Time by Considering Correlation between the Sources Using Different Number of Sources.

Number of Sources	Number of Triples	CPU time	I/o time	Recall	False Positive	Precision
10	562	227,404	220, 227	0.99	0.1519337	0.78260
20	742	1,032,221	997,851	0.99	0.046125	0.887892
30	938	6,273,000	6,264,854	0.97	0.00542	0.97979

2- By using dataset with false rate = 0.2, which means number of false triples is just 20% of the whole data set. We used different number of sources and correctness value to understand the efficiency of the techniques. By using 10, 20, and 30 sources and correctness value as 0.95, we get the results that are shown in the table 29.

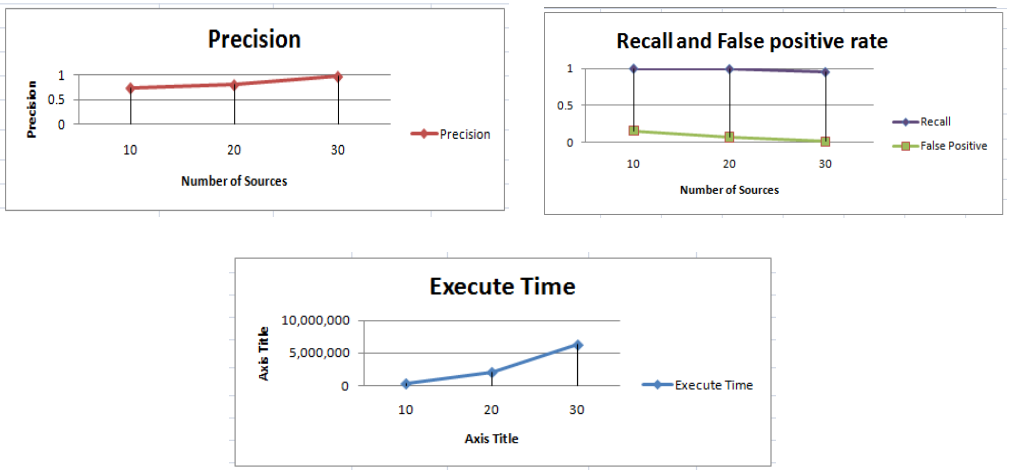

Figure 15. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 20% Random Errors.

Table 29. CPU time by Considering Correlation between the Sources Using Different Number of Sources.

Number of Sources	Number of Triples	CPU time	I/o time	Recall	False Positive	Precision
10	663	294,214	284,730	0.995	0.15550	0.734317
20	927	2,033,955	2,015,533	0.99	0.068775	0.798387
30	1037	6,274,000	6,264,854	0.95	0.007168	0.969387

We calculate the CPU time in each case to show the effectives of the technique. The CPU time includes time it takes to read the data from the excel file, to calculate the correlation between sources, and to find the probabilities of each triple.

6.3.2 Aggressive Approximation

We also did many experiments by considering correlation between the sources but in this time using aggressive approximation method.

First, we show the experiments by using Countries and capitals Dataset (Random False) as it's described in section 5.3. We used different number of source and calculate the precision, recall, and the false positive in each experiment in order to figure out the quality of the technique as its shown below:-

1- By using dataset with false rate = 0.1, which means number of false triples is just 10% of the whole data set. We used different number of sources and correctness value to understand the efficiency of the techniques. By using 10, 20, and 30 sources and correctness value as almost 0.99, we get the results that are shown in the table 30.

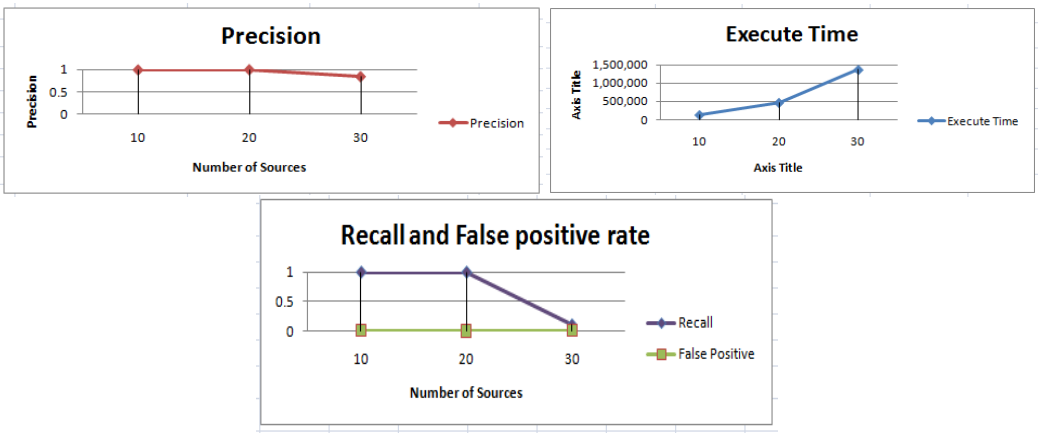

Figure 16. Precision, Recall, False Positive Rate, Execution Time, and the Number of Sources When the Sources Contain 10% Random Errors.

Table 30. CPU Time by Considering Correlation between the Sources Using Different Number of Sources.

Number of Sources	Number of Triples	CPU time	I/o time	Recall	False Positive	Precision
10	562	124,251	95,866	1	0.0055	0.99009
20	742	459,360	430,290	1	0.0018	0.99502
30	938	1,367,319	1,357,710	0.11	0.0054	0.84615

We calculate the CPU time in this case to show the effectives of the technique. The CPU time includes time it takes to read the data from the excel file, to calculate the correlation between sources, and to find the probabilities of each triple.

6.4 Comparing between the Results

It is obvious that a slight difference is there between the results that we get in each case. However, there are large differences when we calculate the initial probabilities with different methods. Consider figure 17 which shows the result of precision by taking five sources and number of false rate=0.2; the initial probabilities in this case is counting by using probabilistic theory approach.

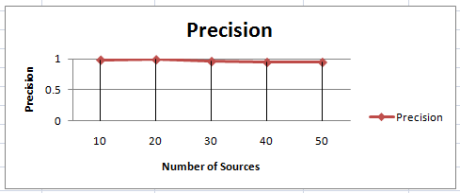

Figure 17. The Precision of 5 Sources with Using the Probabilistic Theory Approach to Compute Initial Probabilities.

Also, consider figure 18 which shows the result of precision by taking five sources and number of false rate=0.2; the initial probabilities in this case is counting by using an ad-hoc approach.

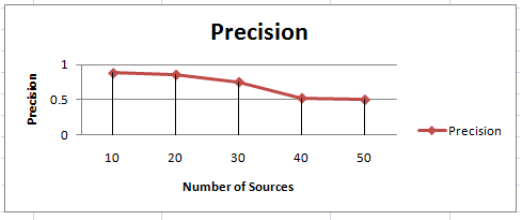

Figure 18. The Precision of 5 Sources with Using an Ad-hoc to Compute Initial Probabilities.

The difference is obvious from the two figures that using probabilistic theory approach to compute the initial probability can get more accurate results.

CHAPTER VII

CONCLUSION AND FUTURE WORK

7.1 Conclusion

We presented three techniques to fuse data; one with not considering correlation between the sources, independent sources, and two with considering correlation between the sources, exact solution and aggressive approximation. These approaches do not require a training set; an initial training set can be obtained using the confidence measures. If a training set is available, the system can use it for improved accuracy. We mentioned the important role of correctness threshold in the fusion process, and presented a method to compute the threshold based on users assessment of the percentage of correct data. We showed the user-assisted threshold approach can significantly improve the accuracy of data fusion.

We present two methods to create datasets in order to have different datasets to validate the effectiveness of the data fusion techniques. The first method creates datasets with random number of errors and the second with intentional errors. Our fusion accuracy was satisfactory for sources containing up to 50% random errors. For intentional falsification, the data fusion accuracy was satisfactory for sources contains up to 20% falsified data, and could be considered acceptable for up to 30% falsification.

For future work, this work experimented by using a large number of sources and triples. We would like to experiment it with significantly larger and more diverse data sets to establish the performance and accuracy guarantees for the fusion. We think sampling-based techniques combined with trustable data, possibly obtained through crowd-sourcing, can be used to provide accuracy guarantees to large-scale data fusion.

REFERENCES

[1] Sadri, F., “On the foundations of probabilistic information integration”, CIKM '12: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, Oct 2012.

[2] Borhanian, A.D., Sadri, F., “A Compact Representation for Efficient Uncertain-Information Integration”, IDEAS '13, Proceedings of the 17th International Database Engineering & Applications Symposium, pp.122-131.

[3] Agrawal, P., Sarma, A. D., Ullman, J., Widom, J., “Foundations of Uncertain-Data Integration”, Proceedings of the VLDB Endowment, Volume 3 Issue 1-2, pp.1080-1090, 2010.

[4] JDL, Data Fusion Lexicon. Technical Panel For C3, F.E. White, San Diego, Calif, USA, Code 420, 1991.

[5] Ravali P., Anish D. S., Xin L. D., Alexandra M., Divesh S., “Fusing Data with Correlations”. In Proceedings of ACM SIGMOD International Conference on Management of Data, pages 433-444, 2014.

[6] Lyublena A., Thomas J., Christoph K., and Dan O. Fast and simple relational processing of uncertain data. In Proceedings of IEEE, International Conference on Data Engineering, pages 983- 992, 2008.

[7] Lyublena A., Christoph K., and Dan O. 10^{106} worlds and beyond: E_cient representation and processing of incomplete information. In Proceedings of IEEE International Conference on Data Engineering, pages 606 - 615, 2007.

[8] Dong F. C., Rada C., Fereidoon S., Tiia J. S. “Query optimization in information integration”. Acta Informatica, 2013. To appear.

[9] Hearst A.M., “Information Integration”, Trends and Controversies, IEEE Intelligent Systems, Sep/Oct1998, http://www.cs.jyu.fi/ai/vagan/course_papers/Paper_12_III.pdf.

[10] CHARU C. A. “Managing and Mining Uncertain Data“, IBM T. J. Watson Research Center, Hawthorne, NY 10532.

[11] Jian P., Bin J., Xuemin L., Yidong Y., Simon F., “Probabilistic Skylines on Uncertain Data”.

[12] Biao Q., Yuni X., Shan W., Xiaoyong D. “A Novel Bayesian Classification Technique for Uncertain Data”.

[13] Waleed A.A., Alaa K. “Handling Data Uncertainty and Inconsistency Using Multisensor Data Fusion”. 2013.

[14] Bahador K., Alaa K., Fakhreddine O. K., Saiedeh N. R. “Multisensor data fusion: A review of the state-of-the-art” . 2011.

[15] Xin L.D., Laure B.E., Divesh S. “Integrating Conflicting Data: The Role of Source Dependence”. Proceedings of the VLDB Endowment, 2(1): 550- 561, 2009.

[16] Nilesh D., Christopher R., Dan S. “Probabilistic Databases: Diamonds in the Dirt communication of the ACM, 52(7):86 -94, 2009.

[17] Dan S., Dan O., Christopher R., Christoph K. “Probabilistic Databases Synthesis Lectures on Data Management”. May 2011.

[18] Prithviraj S., Amol D., Lise G. “Representing Tuple and Attribute Uncertainty in Probabilistic Databases”.

[19] Agrawal, P., “Incorporating Uncertainty in data Management and Integration”, Stanford University, 2012, http://ilpubs.stanford.edu:8090/1053/

[20] Xin l. D., Evgeniy G., Geremy H., Wilko H., Kevin M., Shaohua S., and Wei Z. “ From data Fusion to knowledge fusion”. Proceedings of the VLDB Endowment, 7(10): 881-892, 2014.

[21] Alban G., Serge A., Amelie M., and Pierre S. Corroborating information from disagreeing views. In proceedings of ACM International Conference on Web Search and Data Mining, pages 131- 140, 2010.

[22] Xiaoxin Y., Jiawei H., and Philip S. Y. “Truth Discovery with Multiple Conflicting Information Providers on the Web”. IEEE Transactions on Knowledge and Data Engineering 20 (6): 796- 808, 2008.

[23] Bo Z., Benjamin I. P. R., Jim G., and Jiawei H. “A Bayesian approach to discovering truth from conflicting sources for data integration”. Proceedings of the VLDB Endowment, 5(6): 550-561, 2012.

[24] Reverb: Open Information Extraction Software Project. http://reverb.cs.washington.edu/.